An Intelligent Fault Diagnosis for Machine Maintenance using Weighted Soft-Voting Rule based Multi-Attention Module with Multi-Scale Information Fusion

Zifei Xu[a,b], Musa Bashir [b *], Wanfu Zhang [a], Yang Yang [a], Xinyu Wang [a], Chun Li [a *]

[a]. School of Energy and Power Engineering, University of Shanghai for Science and Technology, Shanghai 200093，P. R. China

[b]. Department of Mechanical and Marine Engineering, Liverpool Logistics, Offshore and Marine (LOOM) Research Institute, Liverpool John Moores University, Liverpool, Byrom Street, L3 3AF, UK

Abstract:

The ability of engineering systems to process multi-scale information is a crucial requirement in the development of an intelligent fault diagnosis model. This study develops a hybrid multi-scale convolutional neural network model coupled with multi-attention capability (HMS-MACNN) to solve both the inefficient and insufficient extrapolation problems of multi-scale models in fault diagnosis of a system operating in complex environments. The model's capabilities are demonstrated by its ability to capture the rich multi-scale characteristics of a gearbox including time and frequency multi-scale

First author: Dr. Zifei Xu (Z.Xu@ljmu.ac.uk)

Corresponding author: * Dr. Musa Bashir (m.b.bashir@ljmu.ac.uk); *Professor Chun Li (lichunusst@usst.edu.cn / lichunusst@163.com)

information. The capabilities of the Multi-Attention Module, which consists of an adaptive weighted rule and a novel weighted soft-voting rule, are respectively integrated to efficiently consider the contribution of each characteristic with different scales-to-faults at both feature- and decision-levels. The model is validated against experimental gearbox fault results and offers robustness and generalization capability with F1 value that is 27% higher than other existing multi-scale CNN-based models operating in a similar environment. Furthermore, the proposed model offers higher accuracy than other generic models and can accurately assign attention to features with different scales. This offers an excellent generalization performance due to its superior capability in capturing multi-scale information and in fusing advanced features following different fusion strategies by using Multi-Attention Module and the hybrid MS block compared to conventional CNN-based models.

Keyword: Intelligent diagnosis; Prognosis and health management; Convolutional Neural Network; Gearbox maintenance; Multi-scale information fusion; Fault diagnosis

I. INTRODUCTION

Future engineering systems need to be equipped with intelligent capabilities in order to improve their efficiency and reduce operational costs through predictive maintenance. Selection of maintenance methods for an equipment is an important consideration in its design life due to its consequence on reduction of economic loss and safe utilization [1]. Available literatures suggest that the development of an effective intelligent maintenance approach is receiving significant attention because of rapid advancements in information technology (data analytics and internet of things) and its application to maintenance industries [2].

Implementation of Prognosis and Health Management (PHM) capabilities in equipment and

machines based on information science is a vital part of the intelligent maintenance approach in the information era. Fault diagnosis is a complimentary part of the PHM, whose task is to identify the different failure modes of the equipment or machines [3]. In the structural health monitoring of a whole machine, it is necessary to obtain information through multiple sensors in order to corroborate fault diagnosis methods, such as in the monitoring of wind turbine operations. Most of the available studies on corroborated fault diagnosis methods focus on the fusion of decision-making with multisensory technologies [4], [5], [6], [7]. However, when monitoring the structural conditions of fine machine components, the space requirement for installation of multi-sensors not only reduces equipment space but also obscures the fault information measurement from both highly sensitive and insensitive sensors. For instance, the bearing vibration of a wind turbine gearbox has sufficient fault information in its main vibration direction [8], but the fault information included in other degrees of freedom (DOFs) is insensitive to the fault patterns. Therefore, this needs fusion of extra decision methods to cover its shortage to conduct effective fault diagnosis [9]. For the intelligent maintenance model to be capable of improving the performance of the fault diagnosis of a system when using a single sensor, it will not only have to solve the local faults on fine components during its condition monitoring but must also lay the foundation for multi-sensor corroborative diagnosis of faults in large and fine components. Therefore, this study focuses on how to achieve an efficient PHM for local mechanical system.

Recent development in methodologies for PHM are largely based on data-driven approaches, which typically consist of feature extraction and pattern recognition. These approaches mostly rely on accurate extraction of fault features from either time-domain or frequency-domain responses of raw vibration data using signal processing algorithms, like the Empirical Mode Decomposition (EMD),

Local Mean Decomposition (LMD) or Variational Mode Decomposition (VMD) [10]. For example, a gearbox's (taken as a partial component in a machine) main function is the sustenance or transmission of dynamic loads on rotating engineering systems. Gearbox typically comprises of gears and bearings and its operating nature makes it susceptible to failure due to defects or faults throughout its design life. These faults and defects can undermine the operational viability of the entire mechanical system or plant, underscoring the need for an intelligent diagnosis method for early detection of any likelihood of fault or failure. The potentials for deformations in the gear and bearing will result in severe safety issues and financial loss to the operators of the actual engineering systems [11].

In the pattern recognition phase, the features extracted from the signal of a faulty system are processed using a machine learning models, such as the Logistic Regression model, Random Forest (regression trees) or Support Vector Machine (SVM) to diagnose fault on gearbox. However, these methods, especially those based on the two phases, have some fatal disadvantages. First, the extracted features that are fed into the machine learning classifiers heavily rely on the knowledge and experience of the users. The upper boundary of the fault diagnosis accuracy is limited by the effectiveness of the extracted features in the data-driven method. Secondly, the generalization ability of a shallow network-based diagnosis model is relatively poor, resulting in a significant difficulty for time-variant working conditions with strong noise interference. Therefore, an alternative solution is urgently needed to efficiently diagnose faults in an engineering system like the gearbox using the measured gearbox signals.

Deep learning, based on a fusion of feature extraction and classification method, can deal with the identified weaknesses in the traditional intelligent fault diagnosis methods by building a model

with high level of advanced features based on deep networks. A majority of generic deep learning algorithms that are currently being widely used, like the Convolutional Neural Network (CNN) model and Deep Belief Network (DBN) model [12], have been examined successfully in the fault diagnosis of systems [13],[14],[15]. However, it is found that the end-to-end CNN-based model, which uses unfiltered vibration responses, can be more effective than the generic fault identification methods based on vibration images [16],[17]. In addition, most studies associated with the CNN-based fault diagnosis method using unfiltered vibration response as input reveal that the fixed scale CNN-based models for diagnosis did not perform well when they are applied to actual working conditions. This is because the characteristics contained in raw vibration signals cannot be sustained on a certain inherent scale where changes in the environments, sampling frequencies or systems exist. The main factors affecting the performance of a multi-scale deep learning method's application in fault diagnosis method are: 1. Extraction and obtaining sufficient information for fault identification; 2. Availability of reasonable strategies for information fusion.

Consequently, Huang *et al.* developed a model called "multi-scale cascade convolutional neural network", whose key idea was to use filters with different scales to provide more useful information [18]. Their results show that considering multi-scale information is effective in improving the accuracy of diagnosis and distinguishing the types of bearing faults. Furthermore, Liu *et al*. [19] added a residual network module into the multi-scale neural network algorithm in order to improve its performance. Their algorithm was designed by incorporating a multi-scale residual neural network, which improves the model's good performance for motor fault diagnosis. In the above studies, although the locations of the feature fusion are different, the processes of the multi-scale feature extraction are both realized

by changing the different receptive fields by the size of convolution kernels. In order to ensure that a CNN-based model can capture multi-scale information in the widest possible receptive field, Zhao *et al*. [20] used the dilated convolution operation to build a multi-scale CNN. The multi-scale factor acts as the dilation factor of convolution kernel. The results show that the dilated multi-scale convolution neural network has a good generalization in bearing fault diagnosis tasks. However, computational complexity is increased when convolution kernels are applied directly to extract multi-scale information from a raw signal. To solve this problem, Jiang *et al*. [21], used a 1-D vibration response signals as the input data and further designed a diagnosis model using the multi-scale CNN (MS-CNN) to diagnose gearbox faults of a wind turbine. Their multi-scale feature extraction was established by the procedure of multi-scale coarse-grained rather than convolution kernels. The results indicated that the time scale of the MS-CNN model offers a considerable influence on the effect of faults diagnosis of the wind turbine gearbox model. Qiao *et al.* [22] designed an adaptive weighted multi-scale with convolutional neural network model in order to conduct an end-to-end diagnosis for an end-to-end bearing model. Their study show that the model offers a strong ability to distinguish faults and with an adaptive ability for the domain against variable operating conditions. Qiao's model considers the differences in contribution between features. Thus, they adaptively weighed the features of different channels and then fused them together. This essentially introduces self-attention into the channels. However, in the above-mentioned multi-scale studies for diagnosis, although multi-scale features are directly fused together at feature-level, the differences in the contribution of advanced features in probability calculation were not considered. Although Xu *et al.* [23] considered these differences in their contribution to advanced feature extractions in probability calculation, they fused the advanced

features on the feature-level that would probably make the useful information meaningless when fusing these advanced features with different scale levels. On the other hand, using non-continuous windows to sample the multi-scale information causes an exponential decrease of the sub-signal's length, leading to hard maintenance of the model and the loss of some important information. Bo *et al.* [24] used separate convolution with depth attention and point attention to establish the multi-scale model but also ignored the differences in the performance generated by the fusion-level in a model. Bias prediction may occur because only a fully connected layer is used to calculate the patterns' probabilities. Zhang *et al*. [25] consider the multiple patterns fused in feature-level may lack of physical meanings. The authors developed a Particle Swarm Optimization (PSO) based weighted majority voting rule to fuse the diagnosed labels in decision-level. However, the PSO-based algorithm is liable to being interwoven and interrupted by other algorithms.

To address the above-mentioned problems, an intelligent diagnosis model based on a novel weighted soft-voting rule-based Multi-Attention with Hybrid Multi-Scale Convolutional Neural Network architecture (HMS-MACNN) is developed in this study. This new model is aimed at solving fault diagnosis problems as demonstrated by its successful application in diagnosing gearbox faults. This paper offers the following specific contributions:

1) An integrated multi-scale block, named Hybrid Multi-Scale (HMS) block, combined the capabilities of multi-scale coarse-grained procedure and different convolutions to simultaneously capture both time multi-scale and frequency multi-scale information of raw vibration signals.

2) An innovative multi-scale CNN-based network (HMS-MACNN) has been proposed to solve the inherent limitations of current methods used in the diagnosis of faults in complex engineering

system like a gearbox. HMS-MACNN is integrated with multi-scale coarse-grained procedures and dilated convolutions to collectively obtain features from raw vibration signals from a system (gearbox) using multiple scales.

3) The proposed multi-Attention block is applied to evaluate the contributions of the hybrid multi-scale features that follows the multi-scale rank in the HMS block, except for channel attention. Multi-Attention block built with two parts including frequency weighted and fused in feature-level and time multi-scale with advanced features weighted and fused in decision-level. This enhances the algorithm' capabilities in the avoidance of undifferentiated and blind features fusion in existing models in accordance with fusion strategies. The decision process in the multi-Attention block consists of a proposed weighted soft-voting rule and an adaptive weighted rule, in which the weighted scores calculated by the weighted soft-voting rule are optimized by a gradient descent without the intervention of other algorithms.

4) As an end-to-end model, the HMS-MACNN is designed by adopting the advantages offered by the 1-D CNN-based model and then extending such benefits to include ability to obtain damage features and diagnose faults from raw vibration signals without applying any manual modifications.

5) This newly developed method improves efficiency and reliability of fault diagnosis tools by returning no false positive results. The method is validated through experimental simulations of gearbox with faults to examine the performance of the newly developed method. The superiority of the HMS-MACNN is further demonstrated through comparison with results from published studies on multi-scale diagnosis models.

Following the introduction, subsequent sections of the paper are structured as follows.

Development of the HMS-MACNN framework is presented in Section 2, while Experimental data and evaluation index are presented in Section 3. Validation, presentation of results and discussion on the HMS-MACNN model under different working conditions are presented in Section 4. Conclusions are presented in Section 5.

II. PROPOSED METHOD

The main constituents of the newly developed intelligent fault diagnosis method are Hybrid Multi-Scale function; a CNN based model, a Multi-Attention Module which consists of an adaptive weighted rule and the proposed weighted soft-voting rule. These components are integrated together to develop the hybrid Multi-Scale CNN with Multi-Attention Architecture. Details of the mathematical formulation, stand-alone performance and capability and the procedure for integrating them in the novel architecture are presented in the following sections.

A. HYBRID MULTI-SCALE

The current methods used for deep learning model-based studies of fault diagnosis of engineering systems use raw vibration signals as the inputs. The reliance of this approach on the raw signal as the main input of the networks essentially constitute a classification or regression problem rather than a diagnosis. However, these models always perform poorly under complex environments such as those with variable loads and strong noises. This is because the CNN-based models have a fixed sampling scale, which can only capture information from a single scale. Therefore, the effective information for fault diagnosis will not remain in an inherent scale when the operating environment changes [26]. Learning the characteristics of a single scale will lead to poor generalization of a model.

The key to overcoming this shortcoming and improving the generalization capability of a CNN-based model is to enhance the model's capability in capturing multi-scale characteristics. Some studies have proposed multi-scale based models, such as MSCNN and MCCNN, which both of them only consider multi-scale characteristics in time and frequency domains independently [15],[18],[21]. However, both time- and frequency-domains multi-scale characteristics are important for fault detection.

Therefore, this study proposed a hybrid multi-scale block (HMS) to be coupled with a CNN model in order to address the fault diagnosis generalization problem. In addition, this method will improve accuracy, efficiency and reliability of fault diagnosis in noisy environments and for systems under complex health and operating environments. A diagram of the hybrid multi-scale block, which consists of the multi-scale coarse-grained procedures [27] and dilated convolutions [28], is illustrated in Figure 1.

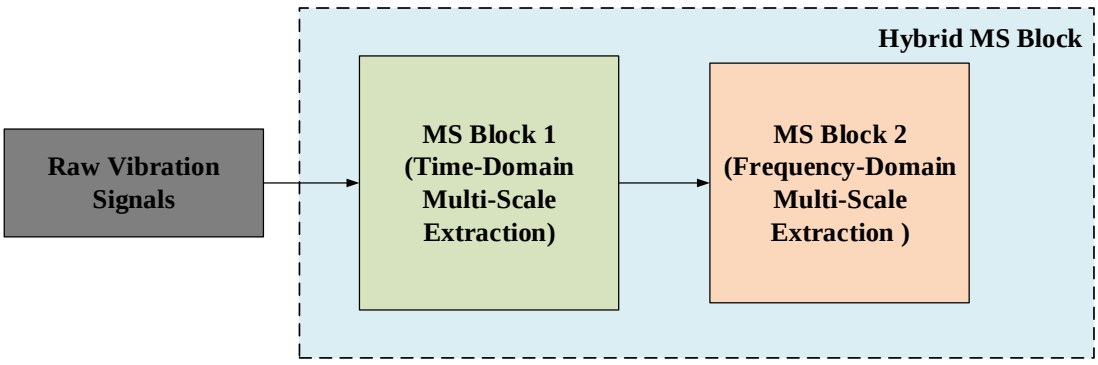


Figure:1 An illustration of Hybrid MS Block

As shown in Figure 1, two multi-scale (MS) blocks are connected in series to form a hybrid MS block. In the hybrid MS block, the MS block1 is used to capture multi-scale features in time domain by using a novel multi-scale coarse-grained procedure. The designed MS block2 is used to extract more useful information in the multi-scale time-frequency domain from the sub-signals that are

obtained by the MS block1. It should be noted that the convolution kernel in MS block2 is of fixed size and the dilated parameters are changed, causing it to have a larger receptive field in a limited amount of data.

Jiang *et al.* [21] used the MS coarse-grained procedure to extract time-domain multi-scale characteristics, but they used the non-continuous window to sample the features, which significantly drops the length of sub-signals and probably misses key information. Thus, MS block1 is designed based on the MS coarse-grained procedure for multi-scale time feature extraction based on improved principles of multi-scale coarse-grained procedure. In addition to being motivated by Zhang *et al.*'s research [29], this study uses a training inference approach in order to improve the MS block1's capability and the model's robustness by using random data point removal technique to augment multi-scale coarse-grained procedure.

The MS block 1 works using a vibration signal in which the response is averaged through a continuous sliding window. In an attempt to improve the designed model's noise resistance, some data points are discarded based on the dropout technique coupled with the procedure of coarse-grained in the training model design phase. The dropout rate p is a selected value set at 0.5.

For each raw vibration signal $x_i : 1 \le i \le n$, the sub-signal. y_j. is obtained by MS block1 at any scale *s* and is calculated by Eq (1).

$$
\begin{aligned}
& r_i^1 \sim Bernoulli(p) \\
& \tilde{x}^l = r_i^1 \cdot x^l \\
& y_j = \begin{cases} \dfrac{1}{\tau} \displaystyle\sum_{i=(j-1)+1}^{(j-1)+1+\tau} \tilde{x}(i),\ j \in [1, n-\tau] \\ y_j = \dfrac{1}{\tau} \tilde{x}(i)\ j \in [n-\tau, n] \end{cases} , \tau \ge 2
\end{aligned}
\tag{1}
$$

where $\tau = 2,3,\ldots,s$ is the factor of scale in MS block1. "$\cdot$" is the element product, r_i^1 follows Bernoulli distribution, which decides whether the data points x^l in the multi-scale coarse-grained procedures $\tilde{x}^l$ has dropped out or not. The length of the sub-signal y_j is n.

In the MS block 2, dilated convolutions with different dilated factors υ are used to extract frequency-domain's multi-scale information. The receptive field calculation of the dilated convolution layer is shown in Eq (2).

$$R = K + (K-1)\cdot(\upsilon-1)\ ,\ \upsilon \in 1,2,\ldots,v \tag{2}$$

Where R is the receptive field of a convolution; K is the kernel size; υ is the dilated factor, which controls the capability of the extraction of frequency-domain multi-scale characteristics.

The hybrid MS block initially obtains s time-domain multi-scale signals using MS Block1. The frequency-domain multi-scale information is extracted by MS Block2 from each multi-scale sub-signal in time domain. Total $s \cdot v$ sub-signals with multi-scale in both time and frequency domains will be obtained by the proposed HMS block.

B. CONVOLUTIONAL NEURAL NETWORK

The feature learning layer consists of parallels of 1D CNNs that extracts representative features from the sub-signals. Generally, the structure of a typical CNN structure comprises of various pairs of convolutional layers, pooling layers, Batch Normalization ((BN)) block and action function. The mathematical process of designing the CNN is given as follows:

$$y^{l(i,j)} = \mathbf{K}_i^l \cdot \mathbf{X}^{l(\mathrm{R}^j)} = \sum_{j'=0}^{W} \mathbf{K}_i^l(j')\mathbf{X}^{l(j+j')} \tag{3}$$

where $\mathbf{K}_i^l$ is the i^{th} filter in the layer l, and $\mathbf{X}^{l(\mathrm{R}^j)}$ is the j^{th} local area of the CNN's layer l;

236 $y^{l(i,j)}$ denotes the dot product of the filter and the Kernel's local area, while W represents the
237 kernel's width. $\mathbf{K}_i^l(j')$ is the j^{th} weight of kernel *l*.

238 ReLU activation function is added to the convolutional layer. The formula for calculating the
239 ReLU function is given in Eq. (4):

$$a^{l(i,j)} = f(z^{l(i,j)}) = \max\{0, z^{l(i,j)}\} \tag{4}$$

240 Where $z^{l(i,j)}$ is the calculated output array of the BN and $a^{l(i,j)}$ is the activation of $z^{l(i,j)}$.

241 In order to efficiently facilitate the neural network training and to overcome gradient
242 disappearance problems caused by activation function, the BN technique is introduced before the
243 pooling operation. The *n*-dimensional array $\mathbf{y}^l = (y^{l(1)}, y^{l(2)}, \ldots, y^{l(n)})$ to the l^{th} BN layer is represented
244 as $\mathbf{y}^{l(i)} = (y^{l(i,1)}, y^{l(i,2)}, \ldots, y^{l(i,n)})$ and $\mathbf{y}^{l(i)} = y^{l(i)} = y^{l(i,1)}$ when the BN layer is positioned
245 immediately after the convolutional layer in the model and followed by the fully connected layer. The
246 BN operation is mathematically represented by the following equations:

$$\hat{y}^{l(i,j)} = \frac{y^{l(i,j)} - \mu}{\sqrt{\sigma^2 + \varepsilon}}, z^{l(i,j)} = \gamma^{l(i)} \hat{y}^{l(i,j)} + \beta^{l(i)} \tag{5}$$

$$\mu = \frac{1}{n} \sum_{i=1}^{n} y^{l(i,j)} \tag{6}$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (y^{l(i,j)} - \mu)^2 \tag{7}$$

247 where $z^{l(i,j)}$ represents the output of a single neuron, while μ and σ^2 are the respective mean and
248 variance of $y^{l(i,j)}$. ε is a negligible normalization quantity (constant) added to sustain the simulation
249 and stops any unexpected termination when the variance is 0. The scale factor and shift parameters to
250 be learned from the features are respectively represented by $\gamma^{l(i)}$ and $\beta^{l(i)}$.

251 A pooling layer, also called the down-sampling layer, is added to HMS-MACNN. This is based

on the most common pooling techniques that includes both average and maximum pooling. However, this research chooses the maximum pooling, and it is presented in Eq (8).

$$p^{l(i,j)} = \max_{(j-1)W+1 \le t \le jW}^{\{a^{l(i,t)}\}} \tag{8}$$

where $a^{l(i,t)}$ represents the value of the t^{th} neuron in the i^{th} framework of the sampling layer l; The corresponding value of the neuron in layer l of the pooling is represented as $p^{l(i,j)}$, and $t \in [(j-1)W+1, jW]$.

The probability distribution of the representative features extracted by CNN are fused and driven into the connected classification layer. Each output is mapped into a probability by a softmax function φ, which is define by:

$$\varphi(u_c) = e^{u_c} \Big/ \sum_{c=1}^{T} e^{u_c}, c = 1, 2, \ldots, T \tag{9}$$

where $\varphi(u_c)$ is a *T*-dimensional probability vector and denotes the probability distribution under *T* kinds of test scenarios, u_c is the output from the CNNs.

C. MULTI-ATTENTION MODULE FOR HYBRID MULTI-SCALE

The multi-scale features have different degrees of sensitivities to failure of gears and bearings when operating in complex environments. However, the existing multi-scale fault diagnosis model indiscriminately integrates the features, which are detrimental to the performance of the models. Although there have been some reported studies on using adaptive weights to evaluate the contributions of features, they essentially used channel attention to evaluate features without considering the differences from contribution of multi-scale characteristics [30],[31].

An adaptive weighted rule, named Attention1 in Multi-Attention Module, is introduced as part of the integration of the hybrid multi-scale with CNN to adaptively score and rank learned features at the

full scales. The implementation of the multi-attention also means that it can assign weights to extracted features learned [32]. Consequently, MS blocks are designed to provide the basis for leaning from the extracted features. The proposed Hybrid MS block consists of two multi-scale blocks. Thus, the features learned by CNNs, which are extracted by MS block1 and MS block2, are weighted through the attention mechanism in turns.

The features $\mathbf{O}_{S\cdot D}: O_1,\ldots,O_i,\ldots,O_{S\cdot D}$ learned from a segment of the raw signals are extracted by HMS Block. A function $G(\cdot)$ has been added to them to obtain features $\mathbf{H}_{S\cdot D}: H_1,\ldots,H_i,\ldots,H_{S\cdot D}$.

$$H_{S\cdot D} = G(O_{S\cdot D}) = \sum_{i=1}^{M} O_{S\cdot D}(i) \tag{9}$$

where $O_{S\cdot D}$ is the i^{th} output feature O_i, and M is obtained using convolution kernels' number obtained from the preceding convolution operation.

An attention module's weights of extracted features on frequency-domain scale $\alpha_{s\cdot 1},\ldots,\alpha_{s\cdot d},\ldots,\alpha_{s\cdot D}$ are obtained by using a fully connected layer coupled with a Softmax function and they are calculated using Eq. (10).

$$\begin{cases} \alpha_{s\cdot d} = \text{Softmax}(\varphi_{s\cdot d}) = e^{\varphi s\cdot d} \Big/ \sum_{\text{d}=1}^{\text{D}} e^{\varphi s\cdot d} \\ \sum_{d=1}^{\text{D}} \alpha_{s\cdot d} = 1 \end{cases} \tag{10}$$

where $\varphi_{s\cdot d}$ is the fully connected layer's output. The weight of an extracted feature of each scale $\alpha_{s\cdot d}$ is calculated using the Softmax function and they are mapped on a probability Space (0, 1), adding to its intelligent capabilities.

The fusion of extracted features Z of MS block2 relating to α_{k} and O_{k} is calculated by Eq.

288 (11).

$$\mathbf{Z}_s = \sum_{d=1}^{D} \alpha_{s\cdot d} O_{s\cdot d}, s = 1, 2, \ldots, S \tag{11}$$

289 Where, $\mathbf{Z}_s$ is the weighted features, which is fused by the frequency-domain multi-scale
290 characteristics.

291 The hybrid multi-scale block is used in the algorithm for the extraction of multi-scale features
292 based on time and frequency in turns. The multi-attention module is used to evaluate the contributions
293 of each advanced features of different scales for fault diagnosis. Different information fusion strategies
294 are used to fuse the time advanced features and frequency features. Thus, a novel weighted soft voting
295 method is proposed to fuse the advanced features at decision level. The weights of soft voting
296 procedure are updated by gradient descent with the network parameters updating.

297 *D. WEIGHTED SOFT-VOTING METHOD FOR ATTENTION ON DECISION-LEVEL*

298 The advanced features, with time a multi-scale fused at feature-level, will typically lack any
299 physical meaning. Thus, the proposed weighted soft-voting rule, named Attention2 in the Multi-
300 Attention Module, fuses the diagnosed results of the time multi-scale advanced features in decision-
301 making level.

302 The probability, $\mathbf{p} = p_1, \ldots, p_2, \ldots, p_s$ is calculated using a fully connected layer and normalized
303 by softmax function corresponding to features $\mathbf{Z}_s$. The final weighted probabilities are calculated by
304 Eq. (12)

$$\mathbf{p}_{weighted} = \boldsymbol{\alpha} \cdot \mathbf{p} = \sum_{s=1}^{S} \alpha_s^c \cdot p_s \,/\, \sum_{s=1}^{S} \alpha_s^c \tag{12}$$

Where $\mathbf{p}_{weighted}$ is the weighted probabilities for fault recognition; $\boldsymbol{\alpha}$ is the weight matrix for soft voting fusion, which is solved by Gradient descent optimization; α_s^c is a weight vector with dimension *c*, and *c* is the total categories.

The loss function of the HMS-MACNN model is a cross entropy between the output probability distributions and predicted categories. The model uses $\mathbf{p}_{weighted}(x)$ as the predicted distribution calculated by weighted soft-voting; $\mathbf{p}(x)$ is the predicted probability distribution that is fused directly, and $\mathbf{q}(x)$ denotes the target probability distribution. The loss between $\mathbf{p}_{weighted}(x)$ and $\mathbf{q}(x)$ is given by Eq. (13), while the loss between $\mathbf{p}(x)$ and $\mathbf{q}(x)$ is given by Eq. (14). Considering the loss between $\mathbf{p}_{weighted}(x)$ and $\mathbf{p}(x)$, the loss function for solving parameters in weighted soft-voting rule is given by Eq. (15)

The loss used to solve the parameters in the weighted soft-voting rule consists of

$$loss_w = -\sum_x p_{weighted}(x) \cdot \log q(x) \tag{13}$$

$$loss = -\sum_x p(x) \cdot \log q(x) \tag{14}$$

$$loss_{weighted} = -\sum_x p_{weighted}(x) \cdot \log q(x) - \sum_x p(x) \cdot \log q(x) + \sum_x p_{weighted}(x) \cdot \log p(x) \tag{15}$$

An illustration of how gradient descent is used to solve weights $\boldsymbol{\alpha}$ in the proposed weighted soft-voting method is shown in Figure 2.

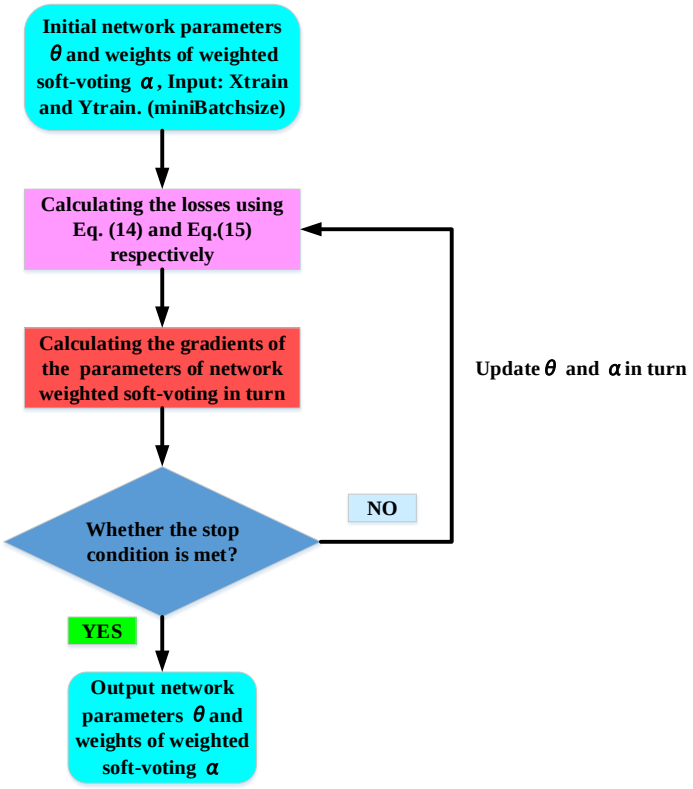


Figure 2: Procedure for solving weights in the weighted soft-voting rule module

As shown in Figure 2, θ is the network's parameter, which is also a parameter in the HMS-MACNN network in addition to parameter α in the weighted soft-voting rule. Two parts of the loss are respectively calculated: a loss not contributing to the weighted soft-voting module; and the other contributing to the weighted soft-voting. Equally, two gradients are calculated separately for each of the losses. Before the iteration stops, the algorithm updates the parameters θ and α based on the two gradients respectively.

E. HYBRID MULTI-SCALE CNN WITH MULTI- ATTENTION ARCHITECTURE

The hybrid model in its basic form, named HMS-MACNN, the frequency multi-scale advance features with attention weights calculated by the Attention MS block 1 are all fused at feature-level. The fused features belong to the advanced features at different time scales .The Attention MS block 2 is used to calculate the weights and then fuses them into the decision-level by the proposed weighted

soft-voting rule. Finally, the weighted probabilities for each classes are calculated for pattern recognition. The framework of the HMS-MACNN is schematically presented in Figure 3.

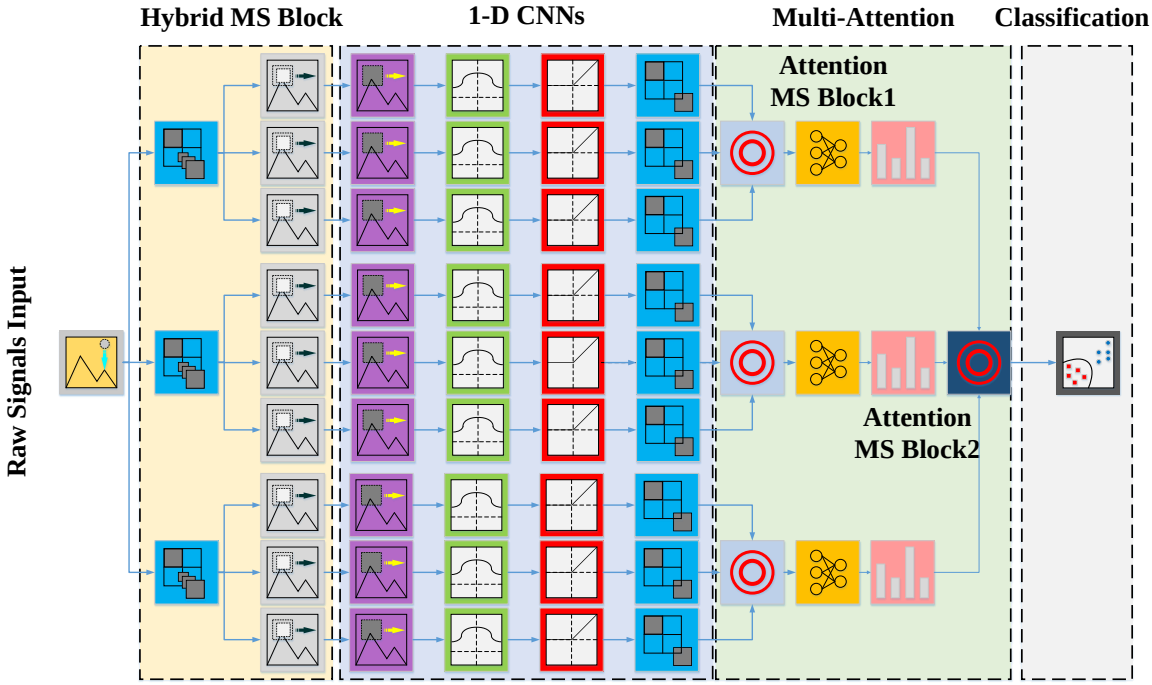


Figure:3 Framework of the HMS-MACNN

In Figure 3, the designed 1-D CNNs block, consisting of several convolution kernel, pooling operation, activation function and BN operation, is used to obtain and learn advanced features of the signals from the multi-scale sub-signals. The weighted features calculated by Eq. (12) are fused in decision-level. Both the scale of MS block1 and MS block2 in the hybrid MS block are taken as 3 in this study. The input size of signals is 4096. Details of the HMS-MACNN architecture are presented in Table I. Note that "CNN" consists of convolution, BN, ReLU and Pooling layers.

Table I: Details of the HMS-MACNN model

No	Layer Name Output size (learnables)	Kernels size /stride Filter number	No	Layer Name Output size (learnables)	Kernels size /stride Filter number
1	Input Layer 4096 (-)	-	2	MS Block 1 4096-3 (-)	(S = 1,2,3)
3	MS Block 2 820-9-16	[128]/[5] 16	4	BN+ReLU+Pool 400-9-16	[3]/[2]

	(2048)	(D = 16,32,64)			
	CNN 1	[3]/[1]		CNN 2	[3]/[1]
5	203-9-32	32	6	100-9-64	64
	(1536)	[3]/[2]		(6144)	[3]/[2]
7	Conv+BN 98-9-128 (24576)	[3]/[1] 128	8	Attention 1 98-3-128	-
9	Fully-Connected (C×37632×3)	-	10	Attention 2 (Classes×3)	-

The fault diagnosis framework is built on the following three steps:

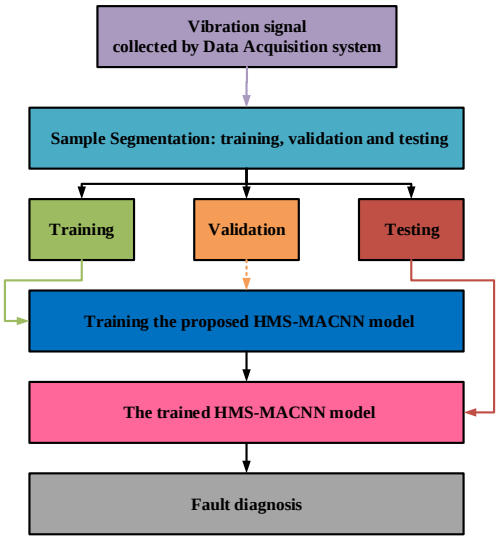


Figure 4：Diagnosis flowchart

Step 1*:* Vibration data from healthy and different operating conditions of a gearbox are collected using a data acquisition system. This is followed by segmentation of the signals into smaller segments for training, validation and testing of the model.

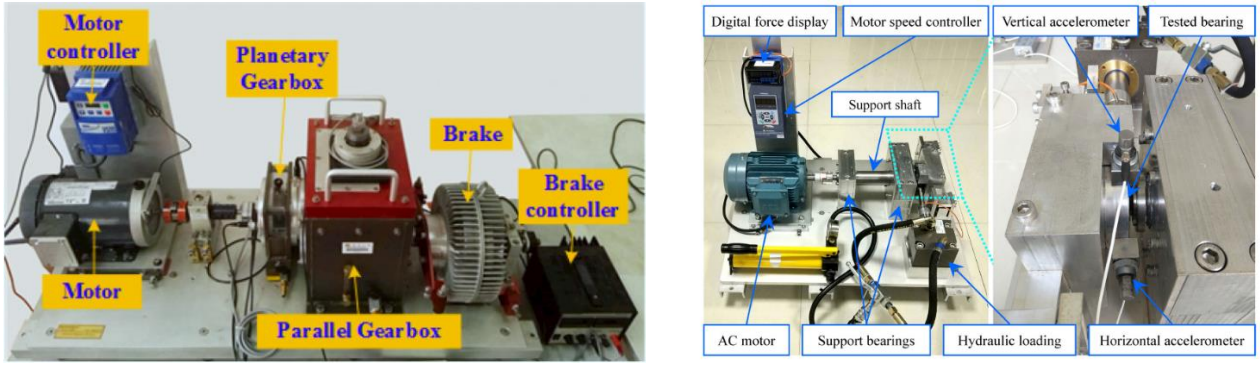
Step 2: developing of end-to-end fault diagnosis system using the training samples obtained from the HMS-MACNN model that were extracted from the raw vibration signals. This step is equipped with an offline training capability. Parameters of the HMS-MACNN model are obtained to prepare for fault diagnosis of the test gearbox following offline training of the input.

Step 3: In this step, testing samples are transferred to the trained HMS-MACNN model for direct diagnosis of the gearbox faults under various scenarios.

III. EXPERIMENTAL DESCRIPTION AND DATA COLLECTION METHOD

In this section, the basic information of the experimental platform and the environment in which the model is examined are introduced. A limited experiment is conducted for the purpose of establishing the credibility of the new method through validation of the results. Two experimental dataset are used in this paper. In one of them, raw vibration signals are obtained from the gearbox (Figure 5(a)) via the drivetrain of the dynamic simulator [33]. In the other dataset, raw vibration signals of bearings acquired by conducting several accelerated degradation experiments are obtained from XI'AN Jiaotong University (XJTU) [34] (Figure 5(b)).

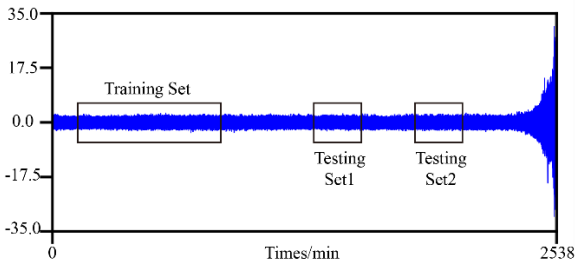
In the gearbox examination, we focused on investigating the two different working conditions in which the loads on the system are set at 20 Hz-0V or 30 Hz-2V. The method used in the study allows for an accurate representation of the health status of the gearbox. Details of bearing and gear types and the respective faults diagnosed are presented in Table II. In the degradation bearing examination, we focused on investigating the diagnosis of both single and mixed faults with degradation. Details of degradation bearings with respective faults are presented in Table III.



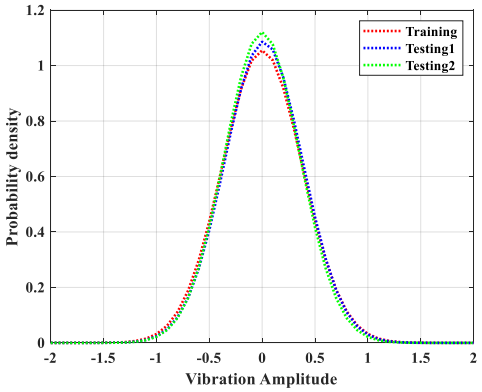
(a) Gearbox rig (b) XJTU bearing degradation experiment rig

Figure 5：Experimental setup

In gearbox fault diagnostic test, experimental dataset from simulations of five different bearing and gear working conditions have been examined in this study. The bearing and gear data comprised of four failure types and one baseline state (healthy condition). Each of the fault types contains 800 training samples and 400 validation samples. The dataset used for the training phase of the bearing and gear comprise of 4096 data points obtained from the input vibration signal. To guarantee high fidelity and ensure consistency across the experiment, the size of dataset used in the testing phase is kept the same as the training dataset. Furthermore, bearing and gear failures are combined into a mixed data set containing the four types of failure stated earlier (four bearing failures, and one baseline [healthy] state) to examine the validity of the newly developed method in diagnosing mixed failures. In the degradation bearing fault diagnostic test, only the bearing failure is considered, but the training and test datasets are established over time. This indicates that the bearing damage of the test dataset may be slightly larger than the training dataset, but belongs to the same type of failure category. The schematic diagram of selecting the training set and the test set is shown in Figure 6.



(a) Dataset chosen



(b) Dataset normal distribution estimation

Figure 6 The dataset segmentation for training and testing

Outcome of the examination confirms the superiority of HMS-MACNN model as demonstrated through comparison of the results obtained in this phase of the study with recently published multi-

382 scale diagnosis studies. This includes comparisons with the MSCNN-I [21], MSCNN-II [20], MCCNN
383 [18] and AWMSC [22].

Table II: Gearbox fault types description

Type	Description	Type	Description
Chipped	Gear feet creak	Miss	Missing gear feet
Root	Gear root feet creak	Surface	Gear surface wear
Ball	Ball creak	Outer	Outer race creak
Inner	Inner race creak	Complex	Inner race and outer race creak

384 The precision of HMS-MACNN in accurately diagnosing faults and failure is further enhanced
385 by the implementation of optimization capability. The Adam and Adadelta gradient descent
386 optimization algorithms are respectively used to optimize the CNN-based model and the weighted
387 soft-voting rule [35], in which are a mini-batch processing size of 200 samples is incorporated in the
388 HMS-MACNN framework. The optimization has 30 epochs in the training phase. Consequently, the
389 learning rate of the feature extraction is initialized to 0.001. This was designed to operate without
390 attenuation during the model updating of each step. A 0.5 dropout rate is adopted for the fully
391 connected CNN layer in order to minimize any exposure of the model to the risk of over-fitting [36].
392 For comparisons, the hyper-parameter settings of each multi-scale model are kept the same as those of
393 HMS-MACNN.

Table III: Bearing degradation description

Type	Description
Inner	Only Inner race damage
Outer	Only Outer race damage
Cage	Only Cage race damage
Hybird	Including Inner race and outer race fault

394 Once again, to guarantee the sustenance of model accuracy during feature learning and
395 classification phases, F1 score is used in the comparison and evaluation of the performance of

diagnosis model examined in this study. This added capability offers the framework a comprehensive metric for measuring its extrapolation function needed in the model updating phase. A mathematical definition of F1 is presented in Eq. (16).

$$F1=\frac{2TP}{2TP+FP+FN} \tag{16}$$

where TP, FP, TN and FN respectively represent the correctly classified faults as positive samples, wrongly classified faults as positive samples, correctly classified faults as negative and wrongly classified faults as negative.

IV. RESULTS AND DISCUSSIONS

Following the successful testing and validation of the framework, the performance of HMS-MACNN model is evaluated using experimental data obtained from the gearbox test rig. The performance is examined under noisy environment and with variable loading conditions.

The comparisons of training and validation accuracies using the proposed HMS-MACNN model and other kinds of multi-scale models examined based on the gearbox dataset are presented in Figure 7. The training and test time over 10 random trails between various MS models are given by Table IV.

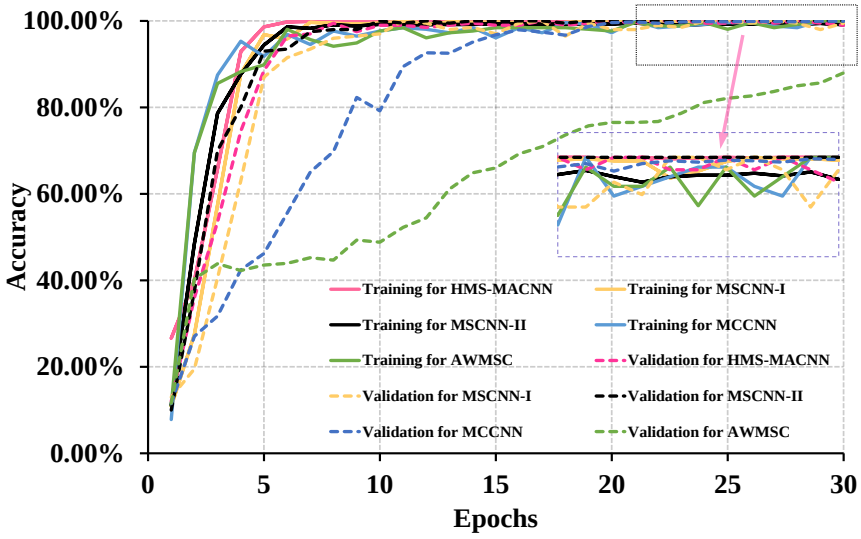


Figure 7: The training and validation accuracy.

409 The training and validation processes of each MS models are presented in Figure 6. Although the
410 proposed HMS-MACNN model's accuracy gradually increases in the first 20 epochs more than those
411 of MSCNN-I and MSCNN-I, the accuracy in the training phase of HMS-MACNN is higher than for
412 other MS models after 20 epochs and equally more stable.

Table IV: COMPARISON BETWEEN VARIOUS MS models FOR OVER 10 RANDOM TRIALS

Methods	Training time (s)	Testing time (s)
HMS-MACNN	397.74±19.61	0.4236±0.0056
MSCNN-I	158.42±6.57	0.3198±0.0055
MSCNN-II	195.27±8.76	0.1352±0.0167
MCCNN	190.13±8.53	0.1609±0.0139
AWMSC	270.10±13.78	0.1464±0.0074

413 Table IV shows that the proposed HMS-MACNN model takes longer time in the training phase
414 due to inherent limitation of computing resources and the process by which the proposed Multi-
415 Attention algorithm handles the dot product of tensors as expected. However, in order to have a
416 realistic basis for comparison with the existing methods, the training process is performed offline and
417 the response time of the proposed HMS-MACNN is completed within a duration of 0.5 seconds only.
418 This slightly exceeds the best duration typically achieved in industrial grade application.

419 In a real industrial application such as wind turbine gearbox that operates in complex operating
420 conditions, the raw vibration signals measured are often drown out by noises. Therefore, the robustness
421 of the HMS-MACNN against noise is examined by injecting additive noise into the signals to
422 reconstruct the raw vibration signals having differences in their signal-to noise ratios (SNR) [37].

423 In this study, the comparison of the robustness of the proposed HMS-MACNN, MSCNN-I,
424 MSCNN-II, MCCNN and AWMSC are examined based on noisy signals with different SNR (-9dB ~
425 9dB). Results from the comparison of the evaluation with the multi-scale models is shown in Figure 8.

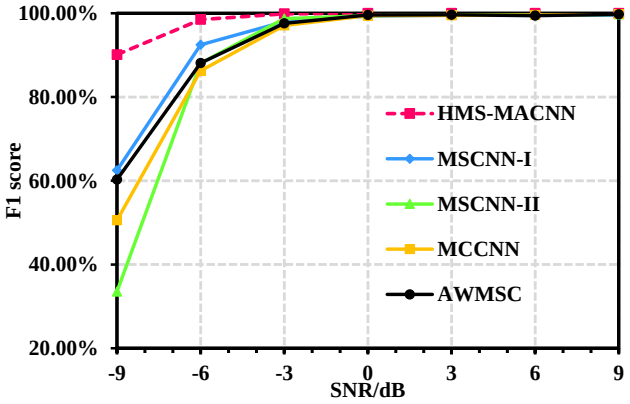
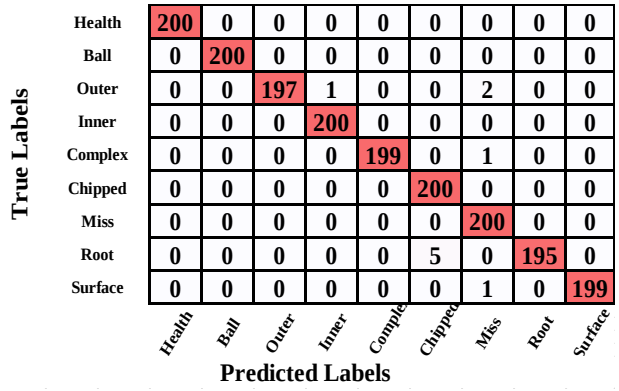


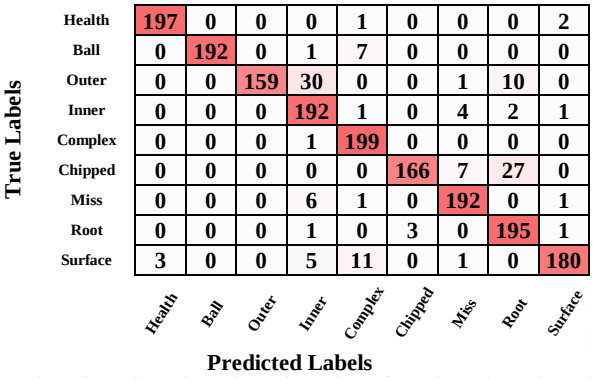
Figure 8: Comparison of anti-noise robustness for model

As shown in Figure 8, when the test signals' SNR is equal to -9dB, the HMS-MACNN model produces an average F1 score of nearly 90%, which are respectively at least 27% higher than other kinds of MS models. This implies that the proposed HMS-MACNN model is more capable of working on data within a noisy environment than the other generic models. Because fusion in decision-level is better than feature-level, the proposed soft-voting rule in the Multi-Attention Module avoids any bias in the prediction. Comparison of performances of the MSCNN-I and MSCNN-II shows that, although MSCNN-II uses the dilated convolution to expand the receptive fields and extract multi-scale characteristics from raw signals, the MSCNN-I performs better than MSCNN-II at -9dB. This is because MSCNN-I has an added multi-scale capability for feature extraction layer that is based on multi-scale coarse-grained procedure. Comparison of the robustness of the MSCNN-I and AWMSC model shows that, although the AWMSC model did not use the coarse-grained method to obtain multi-scale features, they have similar robustness when they face the same strong noisy environments. This is because the AWMSC model focuses more on the features extraction channels. The proposed hybrid MS block is more capable of using the multi-scale characteristics extraction than other generic MS models. The contribution of the features at different scales for different fault types is considered by the HMS-MACNN model because it is capable of giving attention to fault features at different scales. This

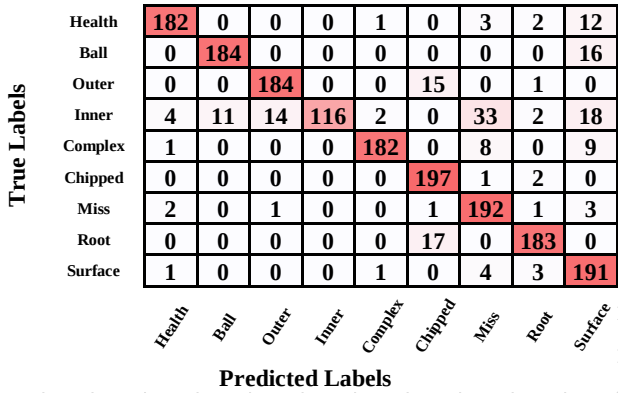
has been demonstrated by the results from the test when the SNR of the test signals increases. The average F1 score of other types of MS models rise quickly and are very close but the HMS-MACNN model still offers the best diagnosis performance. In order to study the misjudgments of fault types by the MS models, the identifications of faults from the MS models at -6dB are shown in Figure 9 using a confusion matrix.



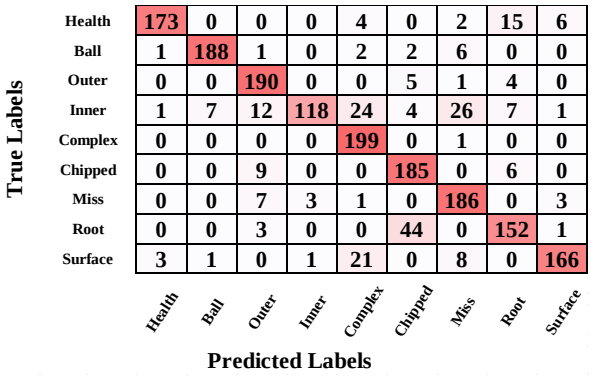
(a) HMS-MACNN



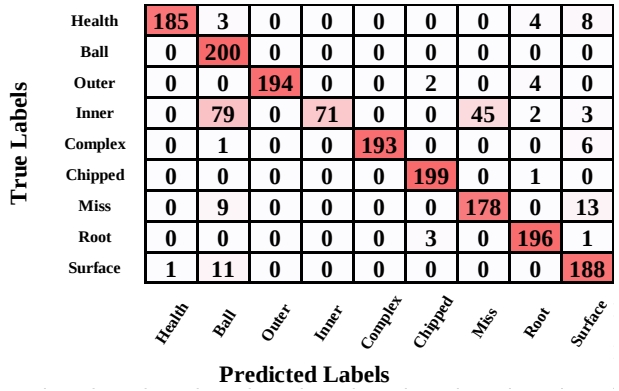
(b) MSCNN-I



(c) MSCNN-II



(d) MCCNN



(e) AWMSC

Figure 9: Comparison through confusion matrix results

As shown in Figure 9, the HMS-MACNN model only has a negligible false positive rate. In the case with a -6dB noise, the positive false that exists between Gear feet creak and outer race creak becomes small due to the proposed weighted soft-voting method. The false alarm rates of the other kinds of MS models are much higher, especially for the inner race fault. Identifying the gear root feet creak under a background with large noise using the MCCNN model is difficult. The performance of the proposed HMS-MACNN model stood out better against many generic MS models. The proposed Multi-Attention block added a key function to the HMS-MACNN that plays a critical role in enhancing the extrapolation capability. The extrapolation capability of a CNN-based model in real world application is important. Therefore, the generalization ability of each MS model is examined by the mixed test set with noise, in which the data of 20Hz-0V accounts for 50% and the data of 30Hz-2V accounts for 50%. A comparison of results with those from other MS models are presented in Table V.

TABLE V: COMPARISON WITH KINDS OF MS MODELS IN TERMS OF F1 SCORE FOR EACH FAULT TYPES (%)

Methods	Health	Ball	Outer	Inner	Complex	Chipped	Miss	Root	Surface	Average
HMS-MACNN	99.34±0.0013	**99.47±0.0033**	**99.46±0.0009**	98.65±0.0036	**99.88±0.0006**	**99.39±0.0056**	98.56±**0.0014**	**98.66±0.0012**	97.33±0.0049	**98.97**
MSCNN-I	95.37±0.0077	98.14±0.0054	89.55±0.0110	91.41±0.0076	95.28±0.0052	88.47±0.0116	93.30±0.0496	92.97±0.0092	94.25±0.0067	93.20
MSCNN-II	91.05±0.0078	90.88±0.0113	87.64±0.0071	81.66±0.0047	85.96±0.0064	91.73±0.0069	89.80.±0.0039	90.52±0.0079	88.03±0.0078	88.59
MCCNN	93.14±0.0093	94.00±0.0054	92.88±0.0060	83.28±0.0110	92.01±0.0064	90.52±0.0088	92.20±0.0611	86.56±0.0128	89.87±0.0087	90.50
AWMSC	96.44±0.0035	87.75±0.0047	97.96±0.0076	78.38±0.0087	97.59±0.0054	96.79±0.0047	87.80±0.0078	95.09±0.0076	90.29±0.0076	92.01

From Table V, the average F1 score shows that the proposed HMS-MACNN model has an accuracy of 98.87% when the testing environments include variable loads and strong noise. This result is 6% higher than the second-ranked method, which is the MSCNN-I model. The comparisons of the F1 score reveal that the HMS-MACNN model successfully performs classification task for nine cases in working conditions (one healthy, eight type of failures). The proposed model has the highest F1 score in identification of six working conditions. The HMS-MACNN model can automatically learn

useful fault characteristics from the raw signals on multiple scales without any manual modifications. This highlights its intelligent capability in diagnosing faults. In addition, the HMS-MACNN model is able to obtain richer information due to its HMS block being assigned with different attention to multi-scale characteristics through the multi-attention block. Therefore, one of the obvious advantages of the proposed HMS-MACNN model is the presence of the multi-scale end-to-end fault diagnosis capability. This uniquely distinguishes the method developed in this study from other generic methods currently being used in the industry.

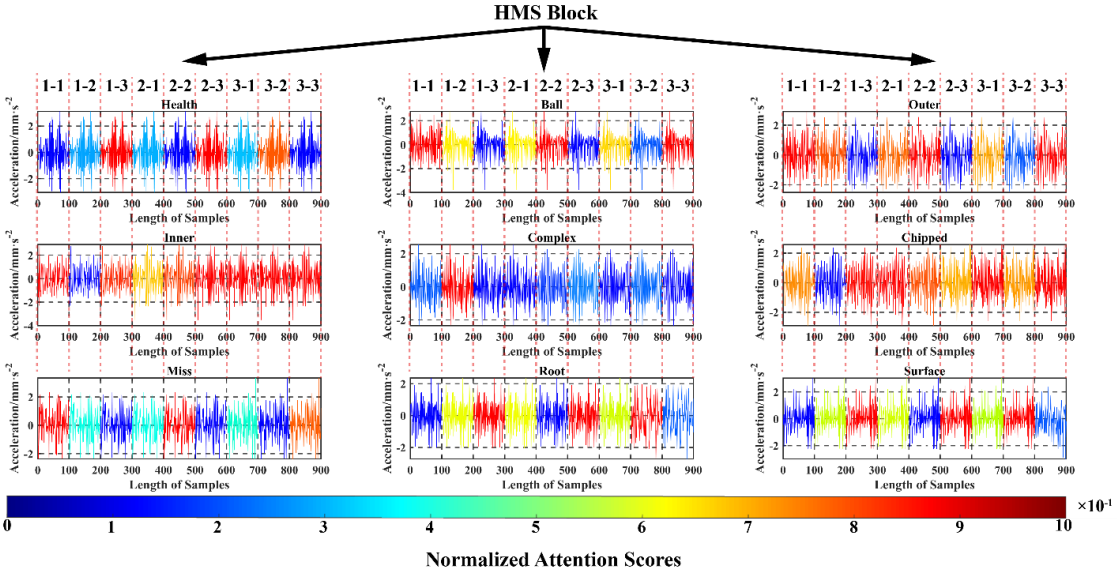


Figure 10: Scores calculated by multi-attention block 1 for a range of gearbox conditions

To further demonstrate the credibility of the proposed HMS-MACNN, the performance of the number of attentions incorporated into the Multi-Attention block for different faults diagnosis are quantified. The attention scores on the multi-time and multi-frequency scale (extracted by HMS block but weighted by Attention 1) for each pattern are expressed in percentage, which are shown in Figure 10. '1-1' means the hybrid multi-scale advanced features with 1 time scale and 1 frequency scale.

Figure 10 shows that the weights given by the multi-attention block1 that the information collected from HMS block with scale = 1-1 is not sensitive to healthy condition, gear surface wear and

gear root creak. The information collected by the HMS block with scale = 1-3 is more sensitive to the health and gear root creak. Each scale has a different sensitivity to different gearbox conditions, which shows the principle of HMS block with multi-attention module to improve the performance of HMS-MACNN.

To further demonstrate the credibility of the proposed HMS-MACNN, the attention scores of weighted self-voting procedure and the confusion matrix of classifications are shown in Figure 11.

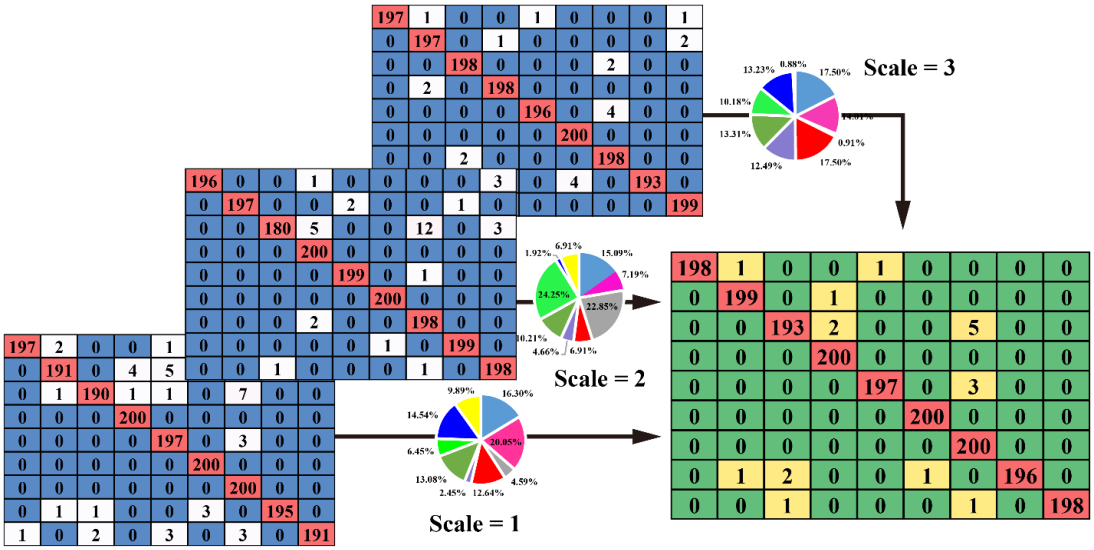
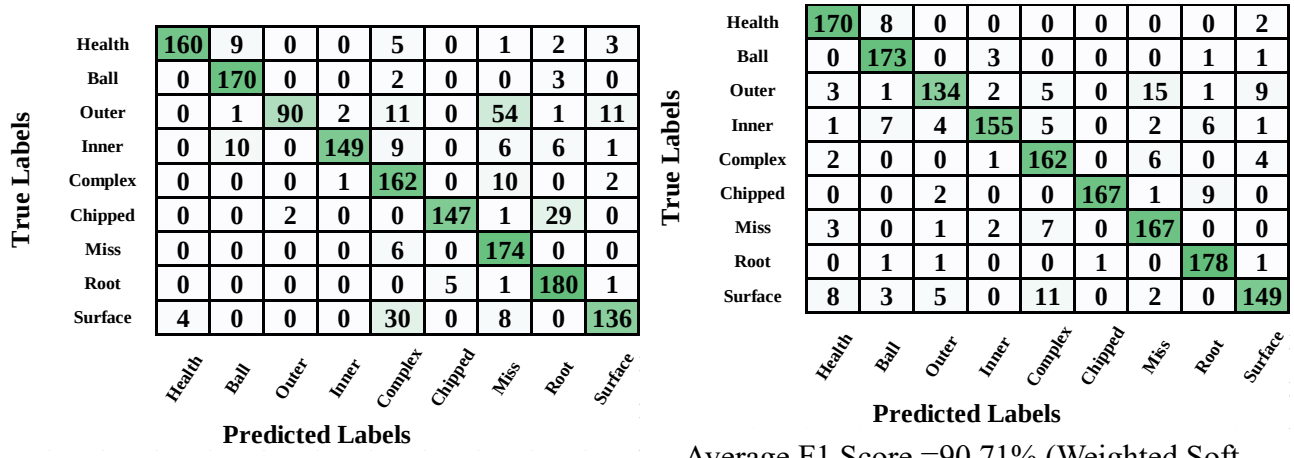


Figure 11: Mechanism of weighted soft-voting rule in Multi-Attention

As shown in Figure 11, scale 1 to scale 3 and final diagnosis results are visualized through the confusion matrix. When the advanced features correspond to scale 1 and scale 2 facing the outer ring fault, the false alarm rate is higher, but scale 3 can recognize the outer race fault better. The false alarm rate in final decision, facing the outer ring fault detection, reduces significantly due to attention put on fusion in the decision level through the proposed weighted soft-voting method.

To further confirm the superiority of the proposed weighted soft-voting in the Multi-Attention module, a comparative assessment based on simulations under large noise scenario with the proposed weighted soft-voting and traditional soft-voting is examined as shown as Figure 12.



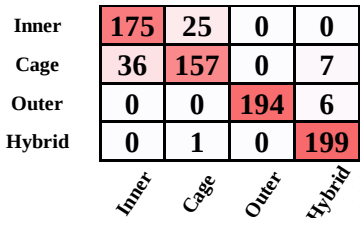
Average F1 Score = 85.03% (Soft-voting)

Average F1 Score =90.71% (Weighted Soft-voting)

Figure 12: Comparison between traditional soft-voting and the proposed weighted soft-voting

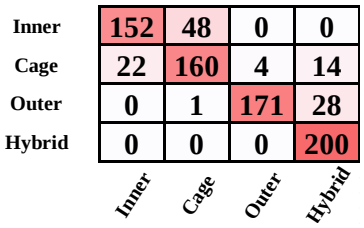
As shown in Figure 12, the proposed weighted soft-voting has an average increase of 5% in F1 score in comparison with traditional soft-voting. Based on the confusion matrix, weighted soft-voting can significantly improve the precision of gear conditions prediction. This further demonstrates the superiority (in terms of efficiency and performance) of the weighted soft-voting rule over a traditional soft-voting. The false positive rates in faults identification are reduced when weighted soft-voting rule is used, further proving its meaningful contribution over traditional soft-voting.

Structural fatigue or improper installation can cause damage to the mechanical structure, but the development of the damage is usually not sudden. Therefore, the model needs to consider the slow changes or evolution of the damage to ensure that the fault can still be identified following any slightest change in material or damage characteristics. Consequently, the degradation test is conducted to prove the extrapolation capability and reliability of the proposed HMS-MACNN model. The results of degradation examination are shown in Figure 13.



(a) Examined by Testing set 1 under -9dB

Average F1 Score =90.50%



(b) Examined by Testing set 2 under -9dB

Average F1 Score =85.31%

Figure 13: Confusion matrix of the test results of HMS-MACNN model examined under different testing set

As shown in Figure 13, the difference between test data 1 and test data 2 is the change in the degree of damage, resulting from the position of the data collection system. The change of the degree of damage will lead to different forms of mechanical vibration, which ultimately affects the data distribution. When using test data 2 to examine the extrapolation of the HMS-MACNN model, its F1 average value is 85.31%, which is 5% lower than Test1's. The reason for this difference is the manner in which samples were collected, with Test2 samples obtained well after training dataset used in Test1 (Figure 6a).

V. CONCLUSION

This study focused on solving the existing problems in fault diagnosis methods for multi-scale systems by developing a framework that is capable of using raw vibration signals from gearbox. A Hybrid Multi-Scale CNN with Multi-Attention (HMS-MACNN) framework for intelligent fault diagnosis of systems such as the gearbox under various operating conditions and different health states is proposed.

The effectiveness of the model is verified via simulations with a gearbox's experimental dataset.

The results confirm that HMS-MACNN is more capable of extrapolating faults from systems operating in a complex environment than the existing multi-scale CNN models considered in this study. Addition of interference in the process of multi-scale time feature extraction can enhance the robustness of the model. The effectiveness of the improved multi-scale block have been have verified using anti-noise tests. The contribution of feature extractions with different scales are calculated using the multi-attention block. The self-attention in multi-attention block 1 effectively gives different high-level feature weights. The weighted soft-voting method in the multi-attention blocks effectively corrects the posterior probability of different scales for pattern recognition of machine conditions, which ultimately reduces the false alarm rates of diagnosis. Therefore, this study addresses the shortcomings of multi-scale fault models developed based on generic CNN models and enhances its efficiency and reliability.

CRediT authorship contribution statement

Zifei Xu: Conceptualization, Methodology, Investigation, Software, Validation, Data curation, Writing – original draft. **Musa Bashir:** Conceptualization, Methodology, Investigation, Resources, Data curation, Supervision, Writing – review & editing, Funding acquisition. **Wanfu Zhang:** Methodology, Writing – original draft, Writing – review & editing. **Yang Yang:** Investigation, Writing – review & editing, Supervision, Funding acquisition, Project administration. **Xinyu Wang:** Conceptualization, Formal analysis. **Chun Li:** Funding acquisition, Supervision.

References

[1] Xu, Z., Li, C., & Yang, Y. . (2020). Fault diagnosis of rolling bearing of wind turbines based on the variational mode decomposition and deep convolutional neural networks. Applied Soft Computing, 95, 106515.

[2] Feng, Z., Liang, M., & Chu, F. (2013). Recent advances in time–frequency analysis methods for machinery fault diagnosis: A review with application examples. Mechanical Systems and Signal Processing, 38(1), 165-205.

[3] Shao, H., Lin, J., Zhang, L., Galar, D., & Kumar, U. (2021). A novel approach of multisensory fusion to collaborative fault diagnosis in maintenance. Information Fusion, 74, 65-76.

[4] Cao, Z., & Chen, L. (2015). Security in application layer of radar sensor networks: detect friends or foe. Security and Communication Networks, 8(16), 2712-2722.

[5] Khaleghi, B., Khamis, A., Karray, F. O., & Razavi, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. Information fusion, 14(1), 28-44.

[6] Al Hage, J., El Najjar, M. E., & Pomorski, D. (2017). Multi-sensor fusion approach with fault detection and exclusion based on the Kullback–Leibler Divergence: Application on collaborative multi-robot system. Information Fusion, 37, 61-76.

[7] Zhang, L., Lin, J., Liu, B., Zhang, Z., Yan, X., & Wei, M. (2019). A review on deep learning applications in prognostics and health management. IEEE Access, 7, 162415-162438.

[8] Li, Y., Jiang, W., Zhang, G., & Shu, L. (2021). Wind turbine fault diagnosis based on transfer learning and convolutional autoencoder with small-scale data. Renewable Energy, 171, 103-115.

[9] Seongpil C, Minjoo C, Zhen G and Togier M. (2021). Fault Detection and diagnosis of a blade pitch system in a floating wind turbine based on Kalman filters and artificial neural network. Renewable Energy, 169：1-13.

[10] Li, Q., & Liang, S. Y. (2018). Weak fault detection for gearboxes using majorization–minimization and

asymmetric convex penalty regularization. Symmetry, 10(7), 243.

[11] Wang, Z., Wang, J., & Wang, Y. (2018). An intelligent diagnosis scheme based on generative adversarial learning deep neural networks and its application to planetary gearbox fault pattern recognition. Neurocomputing, 310, 213-222.

[12] Bengio Y , Courville A , Vincent P . (2013). Representation Learning: A Review and New Perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8):1798-1828.

[13] Wang, J., Zhuang, J., Duan, L., & Cheng, W. (2016). A multi-scale convolution neural network for featureless fault diagnosis. In 2016 International Symposium on Flexible Automation (ISFA) (pp. 65-70). IEEE.

[14] Chen, Z., Gryllias, K., & Li, W. (2019). Mechanical fault diagnosis using convolutional neural networks and extreme learning machine. Mechanical systems and signal processing, 133, 106272.

[15] Jing, L., Zhao, M., Li, P., & Xu, X. (2017). A convolutional neural network based feature learning and fault diagnosis method for the condition monitoring of gearbox. Measurement, 111, 1-10.

[16] Zhao, B., Zhang, X., Li, H., & Yang, Z. (2020). Intelligent fault diagnosis of rolling bearings based on normalized CNN considering data imbalance and variable working conditions. Knowledge-Based Systems, 199, 105971.

[17] Wang, X., Mao, D., & Li, X. (2021). Bearing fault diagnosis based on vibro-acoustic data fusion and 1D-CNN network. Measurement, 173, 108518.

[18] Huang, W., Cheng, J., Yang, Y., & Guo, G. (2019). An improved deep convolutional neural network with multi-scale information for bearing fault diagnosis. Neurocomputing, 359, 77-92.

[19] Liu, R., Wang, F., Yang, B., & Qin, S. J. (2019). Multiscale kernel based residual convolutional neural network for motor fault diagnosis under nonstationary conditions. IEEE Transactions on Industrial Informatics, 16(6), 3797-3806.

[20] Zhao, B., Zhang, X., Zhan, Z., & Pang, S. (2020). Deep multi-scale convolutional transfer learning network: A novel method for intelligent fault diagnosis of rolling bearings under variable working conditions and domains. Neurocomputing, 407, 24-38.

[21] Jiang, G., He, H., Yan, J., & Xie, P.. (2018). Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox. IEEE Transactions on Industrial Electronics, 66(4), 3196-3207.

[22] Qiao, H., Wang, T., Wang, P., Zhang, L., & Xu, M. (2019). An adaptive weighted multiscale convolutional neural network for rotating machinery fault diagnosis under variable operating conditions. IEEE Access, 7, 118954-118964.

[23] Xu, Z., Li, C., & Yang, Y. (2021). Fault diagnosis of rolling bearings using an improved multi-scale convolutional neural network with feature attention mechanism. ISA transactions, 110, 379-393.

[24] Zhao, B., Zhang, X., Zhan, Z., & Wu, Q. (2021). Deep multi-scale separable convolutional network with triple attention mechanism: A novel multi-task domain adaptation method for intelligent fault diagnosis. Expert Systems with Applications, 182, 115087.

[25] Zhang, Z., Han, H., Cui, X., & Fan, Y. (2020). Novel application of multi-model ensemble learning for fault diagnosis in refrigeration systems. Applied Thermal Engineering, 164, 114516.

[26] Misra, M., Yue, H. H., Qin, S. J., & Ling, C. (2002). Multivariate process monitoring and fault diagnosis by multi-scale PCA. Computers & Chemical Engineering, 26(9), 1281-1293.

[27] Zhang, L., Xiong, G., Liu, H., Zou, H., & Guo, W. (2010). Bearing fault diagnosis using multi-scale entropy and adaptive neuro-fuzzy inference. Expert Systems with Applications, 37(8), 6077-6085.

[28] Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122..

[29] Zhang, W., Li, C., Peng, G., Chen, Y., & Zhang, Z. (2018). A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load. Mechanical Systems and Signal Processing, 100, 439-453.

[30] Yang, S., Sun, X., & Chen, D. (2020). Bearing fault diagnosis of two-dimensional improved Att-CNN2D neural network based on Attention mechanism. In 2020 IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS) (pp. 81-85). IEEE.

[31] Zheng, X., Wu, J., & Ye, Z. (2020). An End-To-End CNN-BiLSTM Attention Model for Gearbox Fault Diagnosis. In 2020 IEEE International Conference on Progress in Informatics and Computing (PIC) (pp. 386-390). IEEE.

[32] Xu Z, Mei X, Wang X, Yue M, Jin J, Yang Y, Li C. (2022). Fault diagnosis of wind turbine bearing using a multi-scale convolutional neural network with bidirectional long short term memory and weighted majority voting for multi-sensors, Renewable Energy. Volume 182, 2022, 615-626

[33] Shao, S., McAleer, S., Yan, R., & Baldi, P. (2018). Highly accurate machine fault diagnosis using deep transfer learning. IEEE Transactions on Industrial Informatics, 15(4), 2446-2455.

[34] Biao Wang, Yaguo Lei, Naipeng Li, Ningbo Li, “A Hybrid Prognostics Approach for Estimating Remaining Useful Life of Rolling Element Bearings”, IEEE Transactions on Reliability, pp. 1-12, 2018. DOI: 10.1109/TR.2018.2882682.

[35] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[36] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

[37]Liu, R., Meng, G., Yang, B., Sun, C., & Chen, X. (2016). Dislocated time series convolutional neural architecture: An intelligent fault diagnosis approach for electric machine. IEEE Transactions on Industrial Informatics, 13(3), 1310-1320.