A Novel Ensemble-Learning-Based Convolution Neural Network for Handling Imbalanced Data

Xianbin Wu, Chuanbo Wen*, Zidong Wang, Weibo Liu, and Junjie Yang

***Abstract*—BackgroundBackground: Deep-learning-based fault diagnosis of wind turbine has played a significant role in advancing the renewable energy industry. However, the imbalanced data sampled by the supervisory control and data acquisition systems has led to low diagnosis accuracy. Additionally, deep neural networks can encounter issues like gradient vanishing and insufficient feature learning during backpropagation when the model is too deep. Methods: This article introduces a novel approach that is based on dynamic weight loss functions to modulate unbalanced data and improve diagnostic accuracy by focusing on misclassification of a small sample number. Results: The proposed approach employs a 1D-CNN model and an ensemble-learning-based convolution neural network (EL-CNN) to enhance diversity of models, and complementarity of feature learning. The EL-CNN model addresses the problem of local features being overlooked and provides more accurate results. Conclusions: The effectiveness of this proposed approach is well demonstrated through experimental cases on real wind turbine pitch system fault data. Two different networks using three different loss functions and three state-of-the-art fault diagnosis models are tested, demonstrating the EL-CNN model's superiority.**

***Index Terms*—Fault diagnosis; Deep learning; Imbalanced data; Ensemble learning; Wind turbine; Loss function.**

I. Introduction

The wind power has grown rapidly, resulting in a significant increase of wind turbines (WTs) that are installed. The pitch system, which is responsible for regulating the output power of the fan and controlling the safe braking of the blades, has a fault frequency of up to 35% due to frequent switching of working conditions and mechanical actions. It is crucial to develop effective fault diagnosis and detection means to enhance the reliable operation as well as reduce maintenance costs of WTs. Several studies (e.g., [8], [33], [49]) have emphasized the importance of accurately identifying and promptly addressing faults in the pitch system.

So far, most scholars have focused on studying wind turbine (WT) fault diagnosis based on various signals, including vibration signals [5], [19], acoustic emission (AE) signals [40], lubricating oil parameters [34], and others. For instance, multi-scale convolutional neural networks (CNNs) have been proposed in [19] that leverage fault features in originally sampled vibration signal to diagnose faults in WT gearboxes, exhibiting excellent results in diagnosing faults in critical WT components such as generators [2], blades [42], and gearboxes [22]. However, it is important to note that, all these methods often require installation of some additional sensors, e.g., vibration or acoustic emission sensors, as well as specific data acquisition equipment. Such equipment can be costly and may adversely impact the fault diagnosis of the WT if not installed correctly. With the mature of data-driven fault diagnosis methods and the popularity of WT supervisory control and data acquisition (SCADA) systems, it is quite necessary to develop more convenient and efficient fault diagnosis methods capable of leveraging existing data sources.

Deep learning has found widespread application in various industrial fault diagnosis fields, thanks to its excellent feature extraction and learning capabilities [3], [4], [10], [30], [31], [37], [45], [50]. When dealing with raw fault monitoring data, deep learning is a multi-hidden layer feature-learning approach that easily extracts fault features from available raw data, thereby enabling accurate fault classification [6], [29], [46]. Presently, many scholars have utilized deep neural networks for fan fault diagnosis, obtaining promising results. For instance, in [28], a deep reconstruction WT anomaly detection method has been put forward based on the triplet-convolutional deep autoencoder.

The time scale characteristics of WT monitoring data have recently drawn much attention in the area [39], [43], [44]. For example, in [26], a hybrid model has been proposed for predicting the power of WTs, where the isolated forest algorithm, the synchronous squeeze wavelet transform method, the aquila optimizer, and the long short-term memory network have been integrated together. A new method has been proposed in [21] that combined gated recurrent units (GRUs) with convolutional neural networks (CNNs) by comprehensively considering the spatial and temporal features in the data, and established a model that could effectively detect abnormal data. A threshold self-setting health condition monitoring approach has been developed in [7] based on deep convolutional generative adversarial networks, which has been used for defining a self-setting threshold to monitor WT generator bearings. In [15], the deep residual shrinkage network (DRSN) and the auxiliary classifier generative adversarial network (ACGAN) have been combined to establish a two-layer diagnosis model dedicated to the main bearings of WTs, which could significantly enhance classification accuracy of WT bearings' main faults in noisy environments and have strong diagnostic capabilities in different load states of the data set.

∗ Corresponding author. Email: chuanbowen@163.com

X. Wu and C. Wen are with the College of Electrical Engineering, Shanghai Dianji University, Shanghai 201306, China.

Z. Wang and W. Liu are with the Department of Computer Science, Brunel University London, Uxbridge, Middlesex, UB8 3PH, United Kingdom. (Email: Zidong.Wang@brunel.ac.uk)

J. Yang is with the School of Electric Information Engineering, Shanghai Dianji University, Shanghai 201306, China. (Email: yangjj@sdju.edu.cn)

Currently, there exists a significant imbalance between normal and fault samples in industrial equipment fault diagnosis, which is also prevalent in the data from WT SCADA systems. Since most of the data collected is normal operation data, directly training the model on this data leads to a drastic reduction in its accuracy. Optimization methods to address unsatisfactory classification due to sample imbalance are generally divided into three types: data processing level optimization, algorithm level optimization, and mixed data and hybrid optimization [41]. In the field of deep learning, a special research interest has emerged in combining data processing level optimization with improved loss functions to reduce the adverse classification effects caused by imbalanced data sets [11], [14]. To better utilize the limited fault data, this article proposes a lightweight and effective multi-branch network structure for extracting and learning different features using a new and novel loss function to improve diagnosis capability of faults impacted by imbalanced data sets.

The main contributions are as follows. 1) A new feature learning model is proposed that enhances diversity and complementarity of feature learning through parallel branching. This model can effectively learn features at different levels of the network, address the issue of gradient disappearance during backpropagation, and improve the accuracy of fault classification. 2) A dynamic weight loss function-based modulation strategy for imbalanced data is proposed, which can dynamically adjust the weights of misclassified samples during training based on the prediction probability of the model output. This approach enables the model to adjust the sample weights for different classification difficulties and solves the problem of low classification accuracy in imbalanced datasets. 3) Experiments on a real SCADA WT pitch fault dataset demonstrate method effectiveness in accurately classifying real-world imbalanced datasets. We provide a validated diagnostic model for real WT pitch faults.

Section II provides the theoretical backgrounds and Section III proposes the EL-CNN model and the modulation strategy for imbalanced data. Section IV validates method effectiveness and superiority using real WT pitch fault data. Lastly, Section V concludes this article.

II. Theoretical Background

A. Convolutional Neural Network

A CNN is a hierarchical model comprising deep convolutional layers that take feature representations as inputs and generate a new set of feature representations in the output fusion acceptance domain. Nonlinear functions are used to model the output features, and a linear combination of complex feature representations can be modelled as input. A typical CNN is composed of pooling layers, cross-stacked convolutional layers, and fully connected layers, where the role of a convolutional layer is to extract local features, and different convolution kernels act as various feature extractors. The convolution process is expressed as a filter that extracts features without losing versatility:

$$a_j^l = f\left(\sum_{i=1}^{M} a_i^{l-1} * k_{ij}^l + b_j^l\right) \tag{1}$$

where a_j^l is the jth feature map of the l layer, a_{j-1}^{l-1} is the $(j-1)$th feature map at layer $l-1$, M is number of features of $l-1$ layer, $*$ is a convolution operation, k_{ij} is the convolution kernel weights of the ith feature of layer $l-1$ and the jth feature of layer l , b is the deviation of the jth feature map at the lth layer, $f(\cdot)$ is a nonlinear activation function, k and b are parameters fine-tuned in back propagation at training time.

The input features are downsampled through the pooled inner layer in the pooling layer to reduce data dimensionality and further feature extraction. This process reduces the number of parameters. Typically, pooling operations are implemented using either the maximum pooling function or the average pooling function. The maximum pooling function captures the most salient features of the input data, while the average pooling function holds local information about input data. They can be described respectively as follows:

$$O_m = \max\{C_{j\times k} : (j+1)\times S\} \tag{2}$$

$$O_a = \text{average}\{C_{j\times k} : (j+1)\times S\} \tag{3}$$

where O_m and O_a are the outputs of the pooling layer, and S is step.

The fully connected layer is typically positioned after several pooling or convolutional layers and for subsequent classification, is used to map multidimensional features that are extracted by previous layers to a vector label space. The Softmax function is then used to calculate the relative probabilities between different classes. A classical CNN structure is illustrated in Fig. 1.

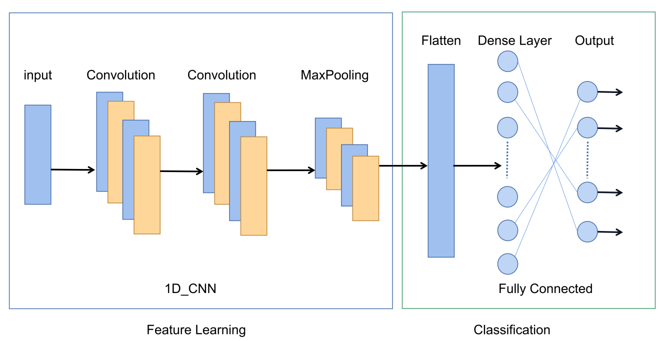


Fig. 1. 1DCNN framework

B. Efficient Channel Attention Mechanism Module

The attention mechanism has been proven to enhance the performance of deep-CNN models effectively. It originated from the field of visual cognitive science, where researchers found that humans pay more attention to target details and suppress otherwise useless information while performing visual tasks such as reading and observing. The attention mechanism in deep learning operates on a similar principle. It aims to make the model focus on crucial features and ignore irrelevant ones, improving the model's convergence speed and accuracy. This technique has achieved remarkable results in object detection, semantic segmentation, and formal explicit [16].

The efficient channel attention (ECA) strategy is a local cross-channel interaction approach that does not require dimensionality reduction and has the effect of attention mechanism on model performance improvement [35]. It also has the advantage of compressing the number of parameters, making it suitable for improving the performance of 1D-CNN models. By applying ECA modules on local cross-channels, the range of interactions can be adaptively adjusted in the module, greatly improving the time cost and computing resource consumption caused by manual parameter optimization. The ECA module structure, as shown in Fig. 2, first obtains accurate features through global average pooling. Then, the channel weight is gained by performing one-dimensional fast convolution of size k, and k can be determined adaptively based on the mapping relationship between k and C. The principle behind this is that there is a correlation between the designated interaction range (i.e., the one-dimensional convolution kernel size K) and the channel dimension C:

$$C = \varphi(k) \tag{4}$$

Assuming that the mapping function is $\varphi(k) = \gamma \times k + b$, however, this linear relationship is not adequate to represent feature number as filter number is typically set to a power of two. To account for this nonlinearity, $\varphi(k)$ can be extended to a nonlinear function

$$C = \varphi(k) = 2^{(\gamma \times k + b)} \tag{5}$$

For a given number of channels C, we can easily obtain an adaptive size of k (5) as follows:

$$k = \phi(C) = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{6}$$

where $|t|_{odd}$ denotes the nearest odd number of t. Clearly, the ECA module allows for a wider range of interactions in high-level channels, whereas low-dimensional channels utilize nonlinear mappings for shorter-range interactions.

III. Proposed Method

A. EL-CNN

In recent years, combining deep convolutional neural networks and ensemble learning has become a hot research topic among scholars [9], [23], [32], [47]. By integrating and filtering the learning results from several different models, the overall accuracy of the decision can be improved, resulting in a more robust classifier. A novel feature-level integration learning approach is proposed that extracts branching feature learning results from an intermediate layer based on the traditional 1D-CNN network of horizontal longitudinal depth. The EL-CNN is then constructed by converging in the final fully connected layer, allowing for multi-level feature fusion to achieve sufficient feature extraction and learning. This enables the model to capture feature information from raw data and learn it through multiple receiver domains with different network structures, providing a reliable and accurate basis for the final classification. Such a network structure offers several benefits as follows.

1) The complexity of the model can be increased without the need to deepen the network architecture, and parallel branching with ECA can better prioritize feature learning without adding more trainable parameters.
2) The introduction of parallel branches allows the leading convolutional layer to be directly connected to the final fully connected layer, effectively resolving gradient disappearance problems during backpropagation due to long network depth.
3) The output of the different branches combines the learning results of multiple receptive fields, effectively solving the problem of too small receptive fields of the convolutional kernels in the deep convolutional layer, allowing for sufficient feature learning.

In each channel of the branch, an ECA module is introduced to enhance accurate feature extraction, and convolutional layers in each branch are expanded with filters of different sizes to increase diversity. Choosing appropriate convolutional kernel sizes from front to back layers, their different sensitivity to features is utilized to achieve a complementary effect. The WT pitch fault data collected by the SCADA system used in this article are all one-dimensional data, and the models are designed to be lightweight to prevent any adverse effect or over-fitting caused by complex models.

The branch structure is shown in Fig. 3. Each branch starts after the pooling layer of the backbone network and uses convolutional layers with different kernel sizes to vary the size of the receptive domain for diversity learning. The ECA module accurately captures important features, and the final output is obtained by integrating multiple features.

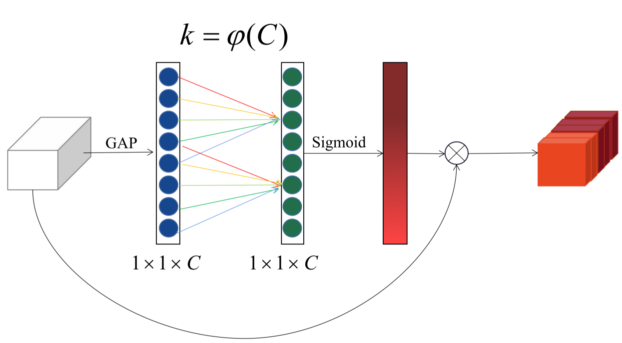


Fig. 2. Structure of ECA module

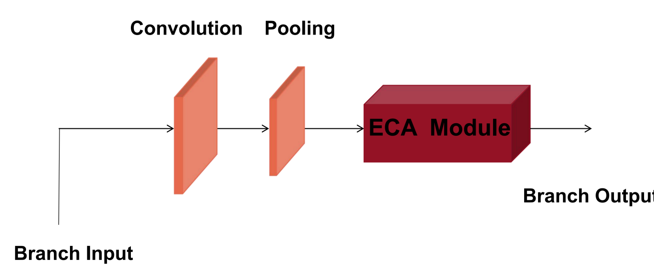


Fig. 3. Structure of Branch

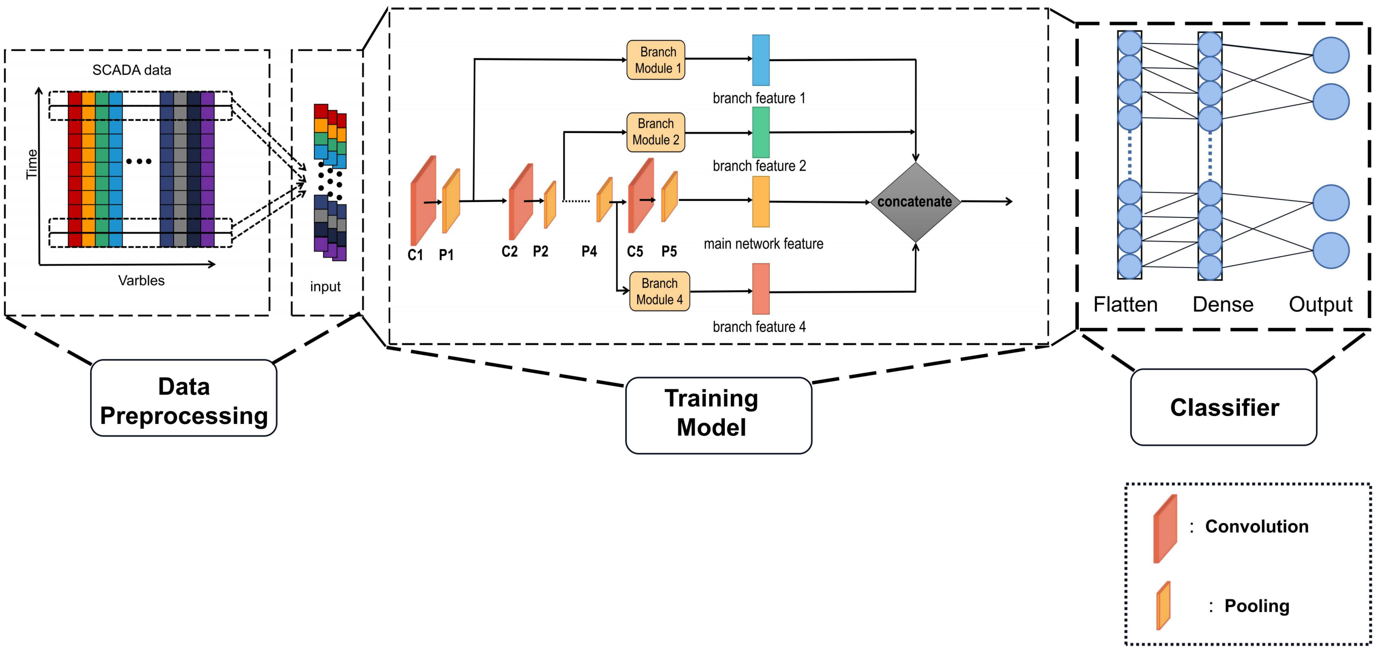


Fig. 4. Structure of the EL-CNN Model

TABLE I. The Network Structure Parameters of EL-CNN

Layer	Operator	Filter	Kernel	Activation	Padding
C1-C5	Conv1D	16;32;64;32;8	1;3;3;2;1	Relu	same
P1-P5	Maxpool1D		2;2;1;2;1		same
BranchC1-C4	Conv1D	8;8;16;16	1;3;5;7	Relu	same
BranchP1-P4	Maxpool1D		2;2;2;2		same

TABLE II. The Parameters of Classifier

Layer	Operator	Size	Activation
D1	Dense	16	Relu
DR2	Drop Out	0.5	/
D2	Dense	32	Relu
D3	Dense	6	Softmax

The overall network architecture is shown in Fig. 4 consisting of a multi-layer CNN network, with each convolutional and pooling layer as a separate sub-module. The branches are connected to the classifier at the end of the network via a Flatten operation. Different branches and different convolutional and pooling modules use different receptive field sizes to ensure that the features are fully learned. The ECA module in the branch models the correlation between the fused feature channels to learn feature weights under different channels, which makes networks more focused on features with valid information.

It should be noted that, in Fig. 4, this model has a backbone and several parallel branches, which can be used to increase the diversity of the model. Each parallel branch outputs features learned on different size receptive domains, allowing the integration of global features while maintaining a focus on local features. Additionally, errors can be back-propagated not only through the backbone network but also through the branches, ensuring that parameters at each network layer are updated in a timely manner and resolving the problem of gradient disappearance when the backbone network is too deep. The EL-CNN model used in this article has five convolutional pooling blocks in its backbone network and four branches.

B. Imbalanced Data Modulation Strategy

1) Review of cross-entropy loss: In multiclass classification problems, the categorical cross-entropy (CE) loss function is commonly used as a measure of classification accuracy error. The training set with n samples is denoted by $D = \{(x_i, y_i)\}_{i=1}^{n} \subset R_x^d \times R_y^d$, where $x \in R^{d_x}$ represents the feature space and $y \in R^{d_y}$ represents the label space. For each data instance, $x_i \in$ X is the input feature vector and $y_i \in Y = \{(1, 2 \cdots, n)\}$ represents the ground-truth class label. The parameter set θ is learned on the map $f : X \to Y$ by minimizing the loss $L(f(x_i, \theta), y)$. Consider a hypothesis (classifier) from a parametric family $F := \{f_\theta : R^{d_x} \times R^{d_y} | \theta \in \Theta\}$. As the Softmax layer is the output layer in the CNN classification model, the categorical CE loss of the CNN classification model can be defined as:

$$L(f(x_i, \theta), y) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{c} y_{ij} \log(f_j(x_i; \theta)) \qquad (7)$$

where θ is the set of parameters of the classifier, y_{ij} is the jth element of one-hot encoded label, $f_j(x_i, \theta) \in R^C$ is the model output with f_j denoting the jth element of f, and n is the sample size. Since the output layer is Softmax, $\sum_{j=1}^{c} f_j(x_i, \theta) = 1$ and $f(x_i, \theta) \geq 0, \forall i, j, \theta$.

2) Dynamic Imbalance Loss: During the training of a CNN model, the weights are updated by backpropagating the error. However, in the case of imbalanced data in multi-classification problems, the distribution of the data across categories may be uneven, causing the classifier to be biased towards the more dominant categories. This bias makes it challenging to classify the few samples that are already difficult to classify accurately. Additionally, as the training process continues, the CE loss can overlook the backpropagation of training errors for minority samples. In industrial settings, it is common for the faulty samples to be much fewer than the normal samples. To address this phenomenon, focal loss (FL) has been proposed to modulate the imbalance phenomenon in the training process. FL is based on CE, with the added effect of reducing the weight of most samples and focusing on training the few samples that are difficult to classify accurately [27]. The FL is defined as follows:

$$L(y, p) = -\alpha(1 - p_t)^{\gamma} \log(p_t) \tag{8}$$

$$p_t = \begin{cases} p & \text{if} \quad y = 1 \\ 1 - p & \text{otherwise} \end{cases} \tag{9}$$

where α is used to modulate the sample weights. In the classification training process, a small number of target categories increase their weights, allowing the misclassified samples to increase their weights. P_t reflects the ease of classification of the sample. The γ modulation can increase the weights of misclassified samples while reducing the weights of correctly classified samples, thereby addressing the challenge of classifying a small number of samples accurately.

Inspired by FL, a class of dynamic disequilibrium loss functions (DL) is proposed in this article, which dynamically adjusts the classification loss based on the classification results. The predicted probability of the classification is used as the basis for dynamic adjustment, while the number of layers of the network model is used as a fixed modulation method. For a multi-classification problem, an ensemble of predicted probabilities for different categories is given by

$$P = \{P_1, P_2, \cdots P_{n-1}, P_n\} \tag{10}$$

The probability that the prediction of class i in a multi-classification passes the Softmax classifier can be expressed as:

$$P_i = \frac{\exp(f_i(x_i; \theta))}{\sum_{j=1}^{n} \exp(f_j(x_i; \theta))} \qquad i = 1, 2 \cdots, n \tag{11}$$

The predicted probability of a model is used as an initial reference to measure classification accuracy. A non-linear function Sigmoid is designed to reflect changes in predicted probabilities, and the results of this mapping are used as a weight for dynamically modulating the sample imbalance error:

$$W_d = \frac{1}{1 + e^{(P_{pred} - 1)}} \tag{12}$$

where $P_{pred} \in P$. Here, the value of P is determined based on the following principles: for a given sample X_i, if its corresponding output label is 1, the corresponding output prediction probability P_1 is used; if the label is 2, then the corresponding output prediction probability is P_2, and so on for all output labels up to n. The prediction probability for class i is calculated using equation (11) after training, and then mapped to dynamic modulation weights using equation (12). The resulting dynamic modulation is reflected in the varying dynamics of the prediction probabilities during each training session, as well as in the probability correlation with other categories.

The choice of the Sigmoid function as the mapping relationship for the predicted weights is made based on two main considerations.

1) The smooth and bounded curve of the Sigmoid function ensures that the weights do not change abruptly during the training process, promoting steady and smooth decrease in the loss value.
2) The fact that Sigmoid mappings are monotonically increasing over the value domain $[0, 1]$ is another important factor in regulating weight reallocation. This ensures that the range of weight variation is not too large.

In addition, a fixed adjustment is also made for branching losses in the EL-CNN model. Together, these adjustments are expressed as follows:

$$W_d = \left(\frac{1}{1 + e^{a \cdot (P_{pred} - 1)}}\right)^{C - \beta + 1} \tag{13}$$

where C represents the number of branches, β represents the number of layers in the current branch, i represents the category of the current classification, and a is an artificially set parameter. This formula assigns more weight to relative predicted probabilities when the predicted probability of a small number of sample outputs is small, effectively redistributing results across different predicted values. Additionally, fixed adjustments to the number of layers in the network can equalize loss values and increase the speed of convergence. The loss function with weights attached can be expressed as:

$$L(p, y) = -\frac{1}{NC} \sum_{i=1}^{N} \sum_{\alpha=1}^{C} W_d \log(p(y = y_i \,|x_i\,)) \tag{14}$$

where N represents the sample size, y is the true value of the sample, p is the probability when the input is x_i and the output prediction label is y_i (that is, P_{pred}). The algorithm for dynamic adjustment of losses is outlined in Algorithm 1.

IV. Experiments

In this section, we demonstrate the effectiveness of our method by validating pitch fault data collected by real SCADA. The hyperparameters regarding the focus loss are illustrated as follows: $\alpha = 0.5$ and $\gamma = 2$. The deep learning framework employed is Keras, with TensorFlow serving as the backend engine. The computer processor utilized is Intel(R) Core i5-5200u-CPU@2.20Ghz, and the CNN models are constructed using Python in the Keras environment.

Algorithm 1 Training Algorithm of Dynamic Imbalance Loss

1: Initialize $x_i = [x_i^1, x_i^2, \cdots, x_i^n]^T$ as model input
2: The EL-CNN model is trained to obtain multiple classification prediction probabilities:
$$P_n = \frac{[\exp(f_1(x_i;\theta)), \cdots, \exp(f_c(x_i;\theta))]}{\sum_{j=1}^{c} \exp(f_j(x_i;\theta))}$$
3: The dynamical weights are obtained by substituting the initially obtained prediction probabilities into equation (13)
4: Use equation (14) to calculate the loss $L(P_n, y)$
5: Repeat steps 3-4 until the maximum epoch is reached
6: The final predicted probability P about the set of label vectors $y_i \in Y = \{(1, 2 \cdots, n)\}$ is the optimal result

TABLE III. SCADA Data Variables

No.	Variable Name
1	Generator active power
2	Generator reactive power
3	Grid active power
4	Grid reactive power
5	Paddles set angle
6	Paddle A pitch angle
7	Paddle B pitch angle
8	Paddle C pitch angle
9	Paddle A pitch torque
10	Paddle B pitch torque
11	Paddle C pitch torque

A. Wind Turbine Pitch SCADA Fault Data

We validate the effectiveness of our method using real SCADA data collected from a 4MW DC-driven hydraulic wind turbine at a wind farm in Shanghai, China. The dataset includes various types of fault data for the wind turbine pitch system from July 2022 to January 2023, comprising a total of 18,000 data samples. The data consists of normal operation and fault data within 10 seconds before and after the fault, with an average sampling rate of 10ms. The operational data includes generator active power, generator reactive power, pitch angle, pitch torque, etc., while the status data includes normal operation and fault status codes, reset time stamp records, etc. The specific data names and variables are listed in TABLE III and TABLE IV.

Unfortunately, the raw SCADA data cannot be used for training immediately as the system directly saves the data before and after 10 seconds, and the data recorded after each fault is not all fault data since the operating status returns to normal when the fault is removed. Therefore, using the data recorded before and after 10 seconds directly leads to abnormal variable correlation and poor diagnostic results. After filtering, we retained 3000 sets of normal data and five different types of fault data, each with an imbalanced ratio to the normal data: 842 sets, 1536 sets, 440 sets, 742 sets, and 440 sets. Each data sample contains 11 variables, with 80% used as the training set and 20% as the validation set.

TABLE IV. Types of Operational Faults

No.	Running Status	label
1	Proportional valves error	0
2	Paris blades misaligned error	1
3	Blade specified position error	2
4	Pitch angle deviation error	3
5	Pitch system low pressure error	4
6	Normal	5

Due to the limited amount of collected data, we use overlapping sampling to expand the dataset without altering the original data. The main parameters for overlapping sampling are the window size and stride. Assuming 1000 data points, with a training sample length (sliding window) of 200 and a step size of 1, the number of samples obtained is $(1000 - 200)/1 + 1 = 801$, which is calculated according to the following formula:

$$\text{sample} = \frac{\text{total\_point} - \text{window\_size}}{\text{stride}} + 1 \tag{15}$$

TABLE V. Performance Metric and Calculated Expression

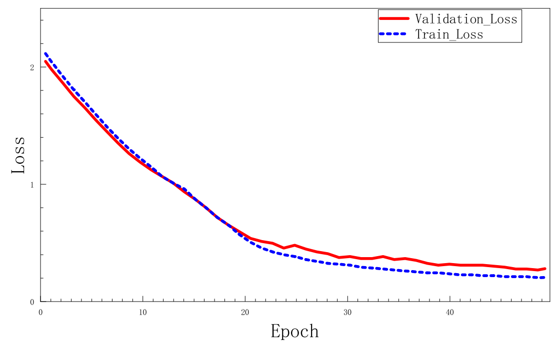
Performance Metric	Calculated Expression
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
f1-score	$\frac{2(\text{precision}\times\text{recall})}{\text{precision}+\text{recall}}$
Micro-Precision(MIP)	$\frac{\sum_{i=1}^{k} TP_i}{\sum_{i=1}^{k}(TP_i+FP_i)}$
Micro-Recall(MIR)	$\frac{\sum_{i=1}^{k} TP_i}{\sum_{i=1}^{k}(TP_i+FN_i)}$
F-score	$\frac{2(MIP\times MIR)}{NIP+MIR}$
Accuracy	$\frac{\sum_{i=1}^{k}(TP_i+TN_i)}{\sum_{i=1}^{k}(TP_i+TN_i+FP_i+FN_i)}$

B. Performance Metric

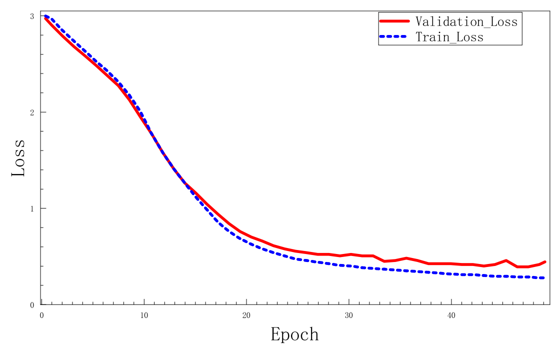
The corresponding evaluation metrics need to be adjusted for the multi-fault classification problem. In addition to accuracy, the F-score is used as an overall classification performance metric as well as a classification metric for individual categories. The calculation formula is given in TABLE V, where TP, TN, FP, and FN refer to true positives, true negatives, false positives, and false negatives in class i, respectively. k represents the number of classes being classified.

C. Contribution of Dynamic Loss to Categorical Performance

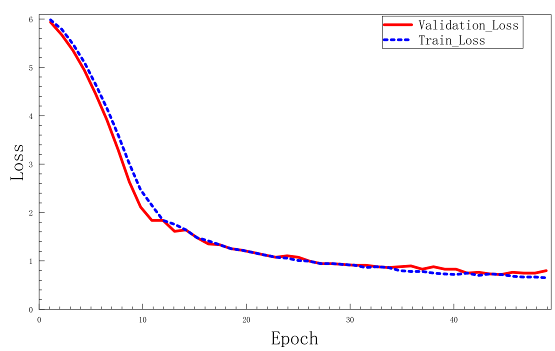
In this subsection, we examine the impact of DL on the performance of fault classification. We compare the loss curves of three models: CNN-CE, CNN-FL, and CNN-DL, which use CE loss, FL, and DL respectively on the CNN model, to validate the improvement effect of DL on the conventional model. We also use the same three loss functions separately on EL-CNN to verify the performance improvement of DL. As shown in Fig. 5, CE loss does not pay much attention to the misclassification of unbalanced data during the training process, resulting in high error from the beginning of training



(a) CNN-DL

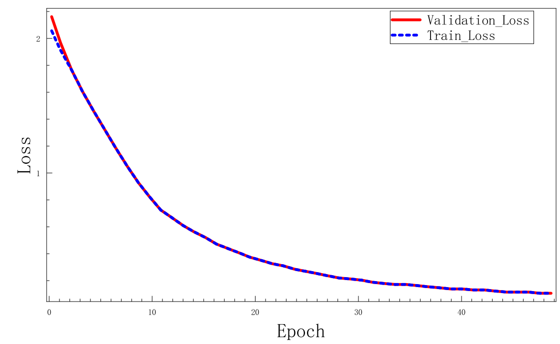


(b) CNN-FL

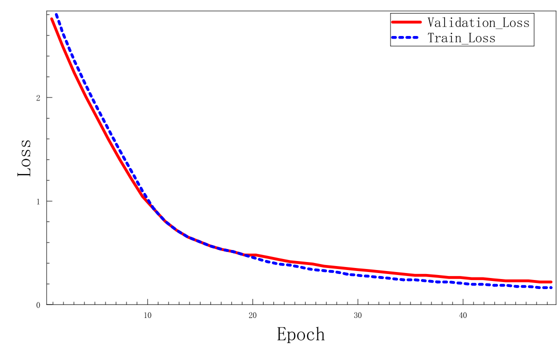


(c) CNN-CE

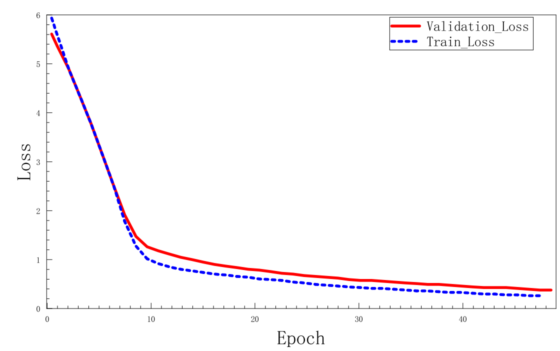
Fig. 5. Training loss and validation loss curves on CNN model



(a) EL-CNN-DL



(b) EL-CNN-FL



(c) EL-CNN-CE

Fig. 6. Training loss and validation loss curves on EL-CNN model

that remains at about 0.8 until the end of the training. However, FL and DL can effectively solve the problem of excessive error. DL uses the predicted value probability of the classification output as a modulation factor to improve the attention to misclassification, and effective weight redistribution of samples can be carried out from the beginning to the end of training to reduce the impact of unbalanced proportion on classification. FL curve and DL curve in Fig. 6 show that DL can be modulated more effectively than FL.

We also verify the effect of different loss functions on the proposed EL-CNN model. The results show that DL has a significant advantage over CE and FL. The low error is maintained throughout the training process, ultimately modulating the loss to a smaller value, providing a good basis for accurate classification and increasing the convergence speed of the whole model on a network of the same magnitude.

Fig. 7 compares the accuracy using three different loss functions on CNN and EL-CNN models. When the models are the same, the dynamic imbalance loss significantly improves the classification accuracy. Through the above analysis, we can prove that the dynamic unbalanced weight modulation strategy has a prominent contribution to the classification of unbalanced data.

D. EL-CNN Contribution to Categorical Results

In this subsection, the effectiveness of the EL-CNN model in fault classification is verified by comparing the classification results of six sets of experimental cases using two evaluation metrics, namely Accuracy and F-score. Additionally, the t-distributed random neighborhood embedding (T-SNE) tech-

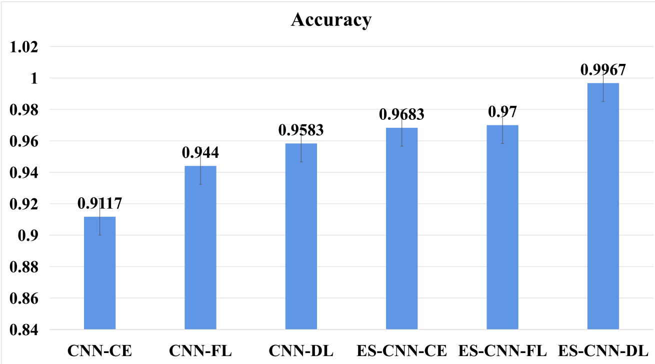
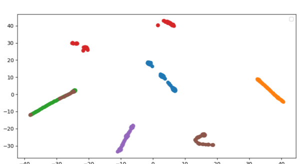


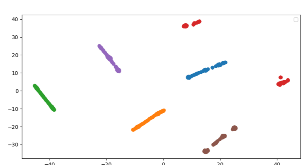
Fig. 7. Comparison of Different Model Accuracy

nique is used to visualize final output distributions, providing a more intuitive representation of the final classification results. In Fig. 8, it can be observed that the change in network architecture has significantly improved the classification of several different faults, and the misclassification phenomenon has been significantly reduced. However, a poor clustering effect is still present in cases (b) and (d). This issue is effectively improved in the scatter plot of classification results from the DL-based EL-CNN model, and this is attributed to that the DL incorporates a computational trick of finding the mean value based on the layer number of networks, improving the situation where the presented clustering effect is not significant due to the increased complexity of the network.

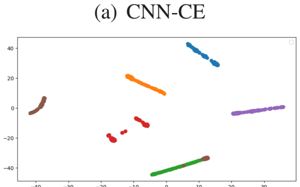
Regarding the training process, Fig. 5 and Fig. 6 showed that around the tenth round of training, the loss using the EL-CNN model was significantly reduced compared to that with the CNN model. As the EL-CNN model solves the problem of gradient disappearance, we find the training and validation losses fall more smoothly and in better agreement when using the same loss function.



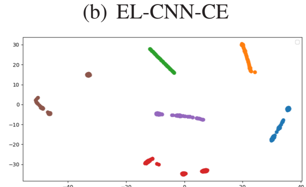
(a) CNN-CE



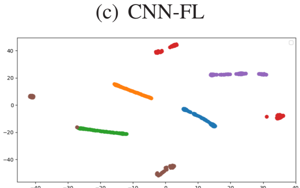
(b) EL-CNN-CE



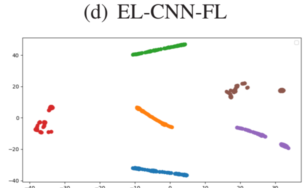
(c) CNN-FL



(d) EL-CNN-FL

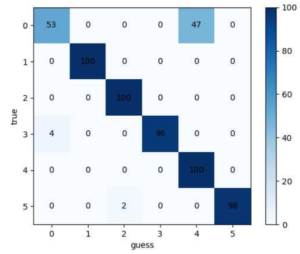


(e) CNN-DL

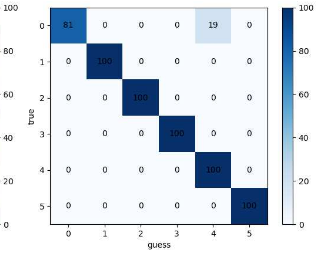


(f) EL-CNN-DL

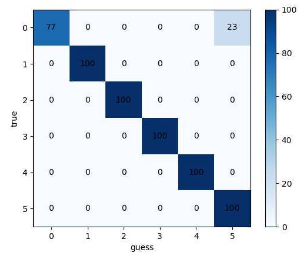
Fig. 8. Scatter plots of different experimental results on a two-dimensional space



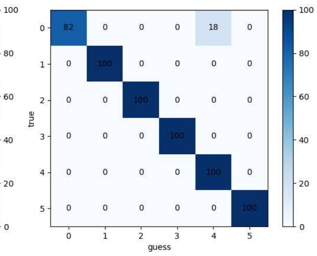
(a) CNN-CE



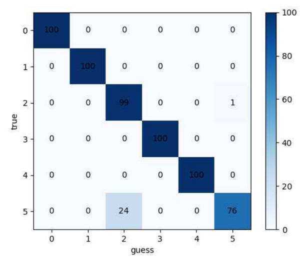
(b) EL-CNN-CE



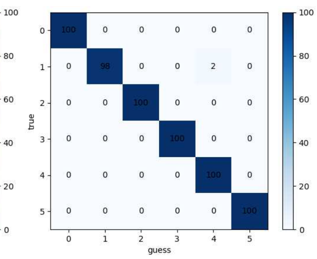
(c) CNN-FL



(d) EL-CNN-FL



(e) CNN-DL



(f) EL-CNN-DL

Fig. 9. Confusion Matrices

In addition, we have introduced a new evaluation metric, the F-score, for classifying unbalanced data. A comparison of the accuracy and F-score of the six experimental sets described above is presented in Fig. 10. From the results, we can observe that the F-score of the EL-CNN model using dynamic imbalance loss is 0.9967, which is significantly higher than the scores of the other five experiments. The confusion matrix for the classification effect is shown in Fig. 9, where the diagonal elements indicate the number of correctly predicted categories. The models that used DL had higher values on the diagonal elements than the results for the remaining four groups, indicating that DL has a significant improvement in the classification of unbalanced data. Moreover, the EL-CNN model with the same loss function showed significant improvements for the misclassified categories. The analysis of the results confirms that our proposed method can significantly enhance the fault classification accuracy of unbalanced fault data.

E. Comparison Experiments with Different Networks

To demonstrate method effectiveness, we conduct comparative experiments with several state-of-the-art 1D-CNN models. During the training process, we use the Adam algorithm to minimize the loss. The following models are evaluated:

(1) MSResnet: a multi-scale residual network based on 1D-CNN, which enhances the model's feature learning capability

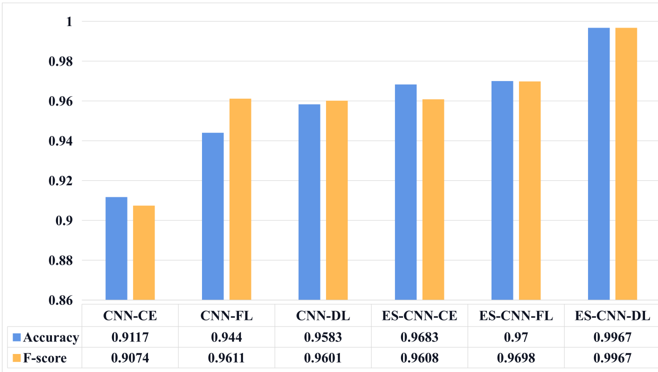


Fig. 10. Comparison of Accuracy and F-score

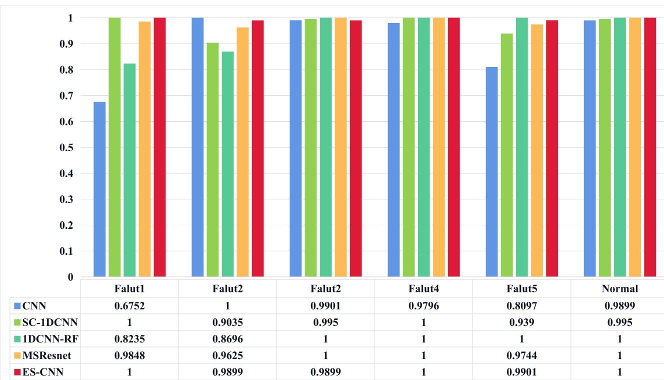


Fig. 11. f1-score of CNN, SC-1DCNN, 1DCNN-RF, MSResnet, and EL-CNN on each fault class

by extracting spatial multi-scale features of SCADA data for WT fault diagnosis [17].

(2) SC-1DCNN: a self-calibrated 1D-CNN based on a fault diagnosis model that can learn shared features from data while considering marginal and conditional distributions [24].

(3) 1DCNN-RF: a 1D-CNN model based on the random forest algorithm, which uses an improved random forest algorithm on the final classifier to optimize the classification results [25].

(4) EL-CNN: the model of this article.

In addition, the conventional 1D-CNN model is utilized as the baseline.

Fig. 11 compares the classification F-scores of the five different models on each type of fault. The EL-CNN model achieves accurate classification on several fault types with a minimum classification score of 0.9899. Compared to other models, it shows more consistent performance in terms of F-score. The overall classification score and accuracy in Fig. 12 indicate that the EL-CNN model outperforms several other methods, achieving significantly better F-score and accuracy. The MSResnet model also shows better classification results, but is less effective than the EL-CNN model in classifying single faults. The 1DCNN-RF model scores higher on the four single fault classifications, but has some shortcomings in scoring lower on the other two fault classifications. The SC-1DCNN model outperforms the traditional CNN model only in fault diagnosis classification using SCADA data single classification score and overall score. Regarding the effect of

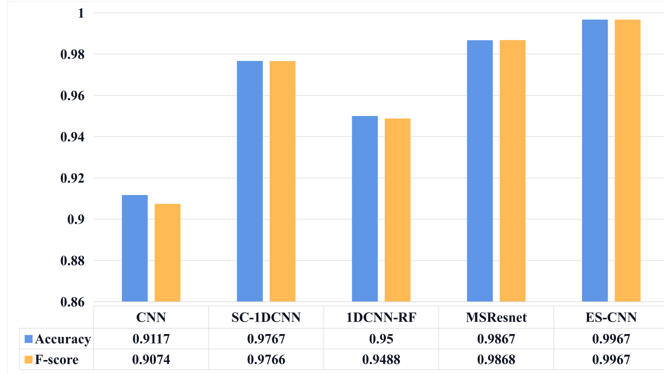
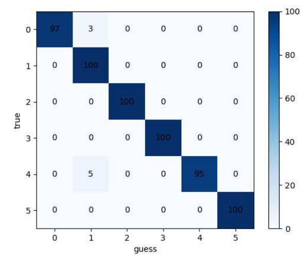
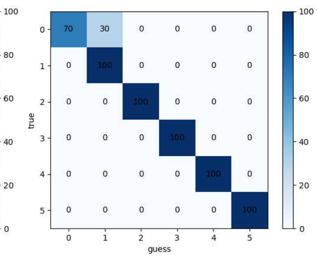


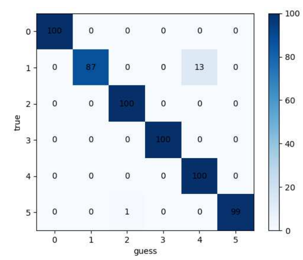
Fig. 12. Overall diagnosis performance of SC-1DCNN, 1DCNN-RF, MSResnet, and EL-CNN on operational SCADA data in terms of Accuracy and F-score



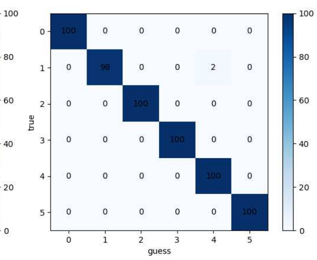
(a) MSResnet



(b) 1DCNN-RF



(c) SC-1DCNN



(d) EL-CNN

Fig. 13. Confusion matrices of different methods.

the diagonal distribution in the confusion matrix in Fig. 13, it is clear that the EL-CNN outperforms several other recent fault diagnosis models in terms of multi-fault classification. Overall, our approach has significant advantages in the application of realistic WT pitch fault data.

V. Conclusion

In this article, a new feature learning model called EL-CNN has been proposed, which enables the learning of local and global features at different network levels to improve model accuracy. The accuracy of fault classification has become better by using the ECA module to accurately capture valid features on different receiver domains. To address the challenge of imbalance in fault data, a dynamic imbalance loss based on the probability of output prediction values has been proposed, which effectively solves the problem of misclassification with small sample size during training. Experimental results comparing with several recent 1D-CNN fault diagnosis models have shown considerable advantages of EL-CNN in fault classification of unbalanced data. Excellent

F-score and Accuracy have been demonstrated in the overall fault classification, and stable F-score values have been shown in each single fault classification. Valid fault features have been extracted from real SCADA WT pitch fault data to accurately classify WT pitch faults, and the results have indicated high practicality for WT pitch fault diagnosis. In the future, we will focus on corresponding research directions, including but not limited to: 1) designing advanced control strategies for fault diagnosis [20], [38]; 2) Applying evolutionary computation methods to automatically tune the hyperparameters [12], [48]; 3) deploying the proposed method to other fault diagnosis and defect detection problems [13], [36]; and 4) developing new fault diagnosis methods for wind turbine pitch [1], [18].

Funding This work was supported in part by the European Union's Horizon 2020 Research and Innovation Programme under Grant 820776 (INTEGRADDE), the Engineering and Physical Sciences Research Council (EPSRC) of the UK, the Royal Society of the UK, the Alexander von Humboldt Foundation of Germany, the BRIEF Award of Brunel University London, the National Natural Science Foundation of China under Grant 61973209, the Natural Science Foundation of Shanghai of China under Grant 20ZR1421200, and the Capacity Building Project of Shanghai Local Colleges and Universities of China under Grant 22010501100.

Data Availability Statement The data that support the findings of this study are available from the corresponding author, C. Wen, upon reasonable request.

VI. Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflicts of interest.

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] W. An, P. Zhao, H. Liu and J. Hu, Distributed multi-step subgradient projection algorithm with adaptive event-triggering protocols: A framework of multiagent systems, *International Journal of Systems Science*, vol. 53, no. 13, pp. 2758–2772, Apr. 2022.

[2] O. Attallah, R. A. Ibrahim and N. E. Zakzouk, CAD system for inter-turn fault diagnosis of offshore wind turbines via multi-CNNs & feature selection, *Renewable Energy*, vol. 203, pp. 870–880, Feb. 2023.

[3] J. Cao, Y. Wang, H. Tao and X. Guo, Sensor-based human activity recognition using graph LSTM and multi-task classification model, *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 18, no. 3s, pp. 1–19, Oct. 2022.

[4] J. Cao, Z. Bu, Y. Wang, H. Yang, J. Jiang and H. Li, Detecting prosumer-community group in smart grids from the multiagent perspective, *IEEE Transactions on Systems Man Cybernetics: Systems*, vol. 49, no. 8, pp. 1652–1664, Aug. 2019.

[5] H. Chen, S. Jing, X. Wang and Z. Wang, Fault diagnosis of wind turbine gearbox based on wavelet neural network, *Journal of Low Frequency Noise Vibration and Active Control*, vol. 37, no. 4, pp. 977–986, Aug. 2018.

[6] J. Chen, W. Hu, D. Cao, Z. Zhang and Z. Chen, A novel multi-task learning method with attention mechanism for wind turbine blades imbalance fault diagnosis, In: *Proceedings of the 2022 4th Asia Energy and Electrical Engineering Symposium (AEEES)*, Chengdu, China, Mar. 2022, pp. 857–862.

[7] P. Chen, Y. Li, K. Wang, M. J. Zuo, P. S. Heyns and S. Baggeröhr, A threshold self-setting condition monitoring scheme for wind turbine generator bearings based on deep convolutional generative adversarial networks, *Measurement*, vol. 117, no. 1, art. no. 108234, Jan. 2021.

[8] A. Dameshghi and M. H. Refan, Wind turbine gearbox condition monitoring and fault diagnosis based on multi-sensor information fusion of SCADA and DSER-PSO-WRVM method, *International Journal of Modelling and Simulation*, vol. 39, no. 1, pp. 48–72, Sept. 2018.

[9] S Dong, W. Feng, Y. Quan, G. Dauphin, L. Gao and M. Xing, Deep ensemble CNN method based on sample expansion for hyperspectral image classification, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, art. no. 5531815, Jun. 2022.

[10] Y. Dong, C. Wen, Z. Wang, A motor bearing fault diagnosis method based on multi-source data and one-dimensional lightweight convolution neural network, *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 237, no. 2, pp. 272–283, Oct. 2022.

[11] J. Du, Y. Zhou, P. Liu, C. -M. Vong and T. Wang, Parameter-free loss for class-imbalanced deep learning in image classification, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 6, pp. 3234–3240, Sept. 2021.

[12] J. Fang, W. Liu, L. Chen, S. Lauria, A. Miron and X. Liu, A survey of algorithms, applications and trends for particle swarm optimization, *International Journal of Network Dynamics and Intelligence*, vol. 2, no. 1, pp. 24–50, Mar. 2023.

[13] J. Fang, Z. Wang, W. Liu, S. Lauria, N. Zeng, C. Prieto, F. Sikström and X. Liu, A new particle swarm optimization algorithm for outlier detection: Industrial data clustering in wire arc additive manufacturing, *IEEE Transactions on Automation Science and Engineering*, in press, DOI: 10.1109/TASE.2022.3230080.

[14] K. R. M. Fernando, C. P. Tsokos, Dynamically weighted balanced loss: Class imbalanced learning and confidence calibration of deep neural networks, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 7, pp. 2940–2951, Jul. 2022.

[15] Z. Fu, Z. Zhou and Y. Yuan Y, Fault Diagnosis of Wind Turbine Main Bearing in the Condition of Noise Based on Generative Adversarial Network, *Processes*, vol. 10, no. 10, Oct. 2022.

[16] A. He, T. Li, N. Li, K. Wang and H. Fu, CABNet: Category attention block for imbalanced diabetic retinopathy grading, *IEEE Transactions on Medical Imaging*, vol. 40, no. 1, pp. 143–153, Jan. 2021.

[17] Q. He, Y. Pang, G. Jiang and P. Xie, A spatio-temporal multiscale neural network approach for wind turbine fault diagnosis with imbalanced SCADA data, *IEEE Transactions on Industrial Informatics*, vol. 17, no. 10, pp. 6875–6884, Oct. 2021.

[18] D. Ji, C. Wang, J. Li and H. Dong, A review: Data driven-based fault diagnosis and RUL prediction of petroleum machinery and equipment, *Systems Science & Control Engineering*, vol. 9, no. 1, pp. 724–747, Oct. 2021.

[19] G. Jiang, H. He, J. Yan and P. Xie, Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox, *IEEE Transactions on Industrial Electronics*, vol. 66, no. 4, pp. 3196–3207, Apr. 2019.

[20] Y. Ju, X. Tian, H. Liu and L. Ma, Fault detection of networked dynamical systems: A survey of trends and techniques, *International Journal of Systems Science*, vol. 52, no. 16, pp. 3390–3409, Nov. 2021.

[21] Z. Kong, B. Tang, L. Deng, W. Liu, Y. Han, Condition monitoring of wind turbines based on spatio-temporal fusion of SCADA data by convolutional neural networks and gated recurrent units, *Renewable Energy*, vol. 146, pp. 760–768, Feb. 2020.

[22] D. Li, Y. Zhao and Y. Zhao, A dynamic-model-based fault diagnosis method for a wind turbine planetary gearbox using a deep learning network, *Protection and Control of Modern Power Systems*,vol. 7, art. no. 22, Jun. 2022.

[23] M. Li, Z. Wang, K. Li, X. Liao, K. Hone and X. Liu, Task allocation on layered multiagent systems: When evolutionary many-objective optimization meets deep Q-learning, *IEEE Transactions on Evolutionary Computation*, vol. 25, no. 5, pp. 842–855, Oct. 2021.

[24] S. Li and J. Yu, Deep transfer network with adaptive joint distribution adaptation: A new process fault diagnosis mode, *IEEE Transactions on Instrumentation and Measurement*, vol. 71, art. no. 3507813, Mar. 2022.

[25] Z. Li, Q. Han, R. Yang, X. Wang and M. Huang, Fault diagnosis for rotating machinery gearbox based on 1DCNN-RF, In: *Proceedings of the 2020 13th International Symposium on Computational Intelligence and Design (ISCID)*, Hangzhou, China, Dec. 2020, pp. 376–379.

[26] Z. Li, Y. Li, X. Luo, M. Liu, X. Cao, S. Du and H. Sun, Short-term prediction of the power of a new wind turbine based on IAO-LSTM, *Energy Reports*, vol. 8, pp. 9025–9037, Nov. 2022.

[27] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, Focal loss for dense object detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[28] J. Liu, G. Yang, X. Li, Q. Wang, Y. He and X. Yang, Wind turbine anomaly detection based on SCADA: A deep autoencoder enhanced by fault instances, *ISA Transactions*, in press, DOI: 10.1016/j.isatra.2023.03.045.

[29] X. Liu, Z. Zhang, F. Meng and Y. Zhang, Fault diagnosis of wind turbine bearings based on CNN and SSAELM, *Journal of Vibration Engineering & Technologies*, DOI: 10.1007/s42417-022-00793-5.

[30] X. Luo, H. Wu and Z. Li, NeuLFT: A novel approach to nonlinear canonical polyadic decomposition on high-dimensional incomplete tensors, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 35, No. 6, pp. 6148–6166, Jun. 2023.

[31] X. Luo, Y. Zhou, Z. Liu and M. C. Zhou, Fast and accurate non-negative latent factor analysis on high-dimensional and sparse matrices in recommender systems, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 35, No. 4, pp. 3897–3911, Apr. 2023.

[32] N. McConnell, A. Miron, Z. Wang and Y. Li, Integrating residual, dense, and inception blocks into the nnUNet, In: *Proceedings of the 2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*, Shenzen, China, Jul. 2022, pp. 217–222.

[33] W. Qiao and D. Lu, A survey on wind turbine condition monitoring and fault diagnosis-Part II: Signals and signal processing methods, *IEEE Transactions on Industrial Electronics*, vol. 62, no. 10, pp. 6546–6557, Feb. 2023.

[34] M. Schlechtingen and I. F. Santos, Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection, *Mechanical Systems and Signal Processing*, vol. 25, no. 5, pp. 1849–1875, Jul. 2011.

[35] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo and Q. Hu, ECA-Net: Efficient channel attention for deep convolutional neural networks, In: *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA, Jun. 2020, pp. 11531–11539, Aug. 2020.

[36] X. Wang, Y. Sun and D. Ding, Adaptive dynamic programming for networked control systems under communication constraints: a survey of trends and techniques, *International Journal of Network Dynamics and Intelligence*, vol. 1, no. 1, pp. 85–98, Dec. 2022.

[37] Z. Wang, C. Wen and Y. Dong, A method for rolling bearing fault diagnosis based on GSC-MDRNN with multi-dimensional input, *Measurement Science and Technology*, vol. 34, art. no. 055901, Feb. 2023.

[38] P. Wen, X. Li, N. Hou and S. Mu, Distributed recursive fault estimation with binary encoding schemes over sensor networks, *Systems Science & Control Engineering*, vol. 10, no. 1, pp. 417–427, Apr. 2022.

[39] M. Wu, T. Wan, X. Wan, Z. Fang and Y. Du, A new localization method for epileptic seizure onset zones based on time-frequency and clustering analysis, *Pattern Recognition*, vol. 111, art. no. 107687, Mar. 2021.

[40] D. Xu, P. F. Liu and Z. P. Chen, Damage mode identification and singular signal detection of composite wind turbine blade using acoustic emission, *Composite Structures*, vol. 255, no. 1, art. no. 112954, Jan. 2021.

[41] F. Yang, D. Huang, D. Li, Y. Zhao, S. Lin and S. M. Muyeen, A novel convolutional network with a self-adaptation high-pass filter for fault diagnosis of wind turbine gearboxes, *Measurement Science and Technology*, vol. 34, no. 2, art. no. 025024, Nov. 2022.

[42] Y. Yu, H. Cao, X. Yan, T. Wang and S. S. Ge, Defect identification of wind turbine blades based on defect semantic features with transfer feature extractor, *Neurocomputing*, vol. 376, pp. 1–9, Feb. 2020.

[43] Y. Yuan, G. Ma, C. Cheng, B. Zhou, H. Zhao, H.-T. Zhang and H. Ding, A general end-to-end diagnosis framework for manufacturing systems, *National Science Review*, vol. 7, no. 2, pp. 418–429, Feb. 2020.

[44] Y. Yuan, X. Tang, W. Zhou, W. Pan, X. Li, H.-T. Zhang, H. Ding and J. Goncalves, Data driven discovery of cyber physical systems, *Nature Communications*, vol. 10, no. 1, pp. 1–9, Oct. 2019.

[45] N. Zeng, P. Wu, Z. Wang, H. Li, W. Liu and X. Liu. A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection, *IEEE Transactions on Instrumentation and Measurement*, vol. 71, art. no. 3507014, Feb. 2022.

[46] R. Zeng and Y. Song, A fast routing capsule network with improved dense blocks, *IEEE Transactions on Industrial Informatics*, vol. 18, no. 7, pp. 4383–4392, Jul. 2022.

[47] K. Zhang, B. Tang, L. Deng and X. Liu, Hydbrid atention improved Resnet based falut daignosis method of wind turbines gearbox, *Measurement*, vol. 179, art. no. 109491, Jul. 2021.

[48] Y. Zhang, L. Chen, Y. Li, X. Zheng, J. Chen and J. Jin, A hybrid approach for remaining useful life prediction of lithium-ion battery with adaptive Levy flight optimized particle filter and long short-term memory network, *Journal of Energy Storage*, vol. 44, art. no. 103245, Dec. 2021.

[49] Y. Zheng, W. Zhou, W. Yang, L. Liu, Y. Liu and Y. Zhang, Multivariate/minor fault diagnosis with severity level based on Bayesian decision theory and multidimensional RBC, *Journal of Process Control*, vol. 101, pp. 68–77, May 2021.

[50] L. Zou, Y. Li and F. Xu, An adversarial denoising convolutional neural network for fault diagnosis of rotating machinery under noisy environment and limited sample size case, *Neurocomputing*, vol. 407, pp. 105–120, Sept. 2020.