IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 12, 2021, accepted October 24, 2021, date of publication October 28, 2021, date of current version November 8, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3124025*

# Fault Diagnosis Methods Based on Machine Learning and Its Applications for Wind Turbines: A Review

**TONGDA SUN, GANG YU, MANG GAO, LULU ZHAO, CHEN BAI, AND WANQIAN YANG**
School of Mechanical Engineering and Automation, Harbin Institute of Technology, Shenzhen, Guangdong 518055, China

Corresponding author: Gang Yu (gangyu@hit.edu.cn)

**ABSTRACT** With the increase in the installed capacity of wind power systems, the fault diagnosis and condition monitoring of wind turbines (WT) has attracted increasing attention. In recent years, machine learning (ML) has played a crucial role as an emerging technology for fault diagnosis in wind power systems has played a crucial role. Even though ML methods have shown great potential in dealing with the issues related to the fault diagnosis of WT, there are still some challenges encountered in many aspects. In this paper, typical fault diagnosis methods based on ML methods for wind power systems are thoroughly reviewed in terms of both theoretical fundamentals and industrial applications, including traditional machine learning (TML), artificial neural networks (ANN), deep learning (DL) and transfer learning (TL), in the development line of ML technologies. The advantages and disadvantages of various methods are analyzed and discussed. Meanwhile, a distribution diagram is provided for the discussions of ML methods applied for WT fault diagnosis, and the existing challenges on the applications for fault diagnosis based on ML for wind power generation systems are presented. Moreover, some prospects for future research directions are provided.

**INDEX TERMS** Wind turbines, machine learning, fault diagnosis, review.

## NOMENCLATURE

| | |
|---|---|
| AE | autoencoder. |
| AI | artificial intelligence. |
| ANN | artificial neural networks. |
| ART | adaptive resonance theory. |
| BPNN | back propagation neural network. |
| CA | clustering algorithm. |
| CNN | convolutional neural network. |
| DAE | denoising autoencoder. |
| DBN | deep belief network. |
| FCM | fuzzy C-means clustering. |
| HMM | hidden Markov model. |
| LSSVM | least squares support vector machine. |
| LSTM | long short-term memory network. |
| ML | machine learning. |
| RBFNN | radial basis function neural network. |
| RF | random forest. |
| RLM | extreme learning machine. |
| RNN | recurrent neural network. |
| RVM | relevance vector machine. |
| SAE | stacked autoencoder. |
| SCADA | supervisory control and data acquisition. |
| SDAE | stacked denoising autoencoders. |
| SOM | self-organizing map. |
| SVM | support vector machine. |
| t-SNE | t-distributed logistic neighbor embedding. |
| TL | transfer learning. |
| TML | traditional machine learning. |
| VMD | variational mode decomposition. |
| WT | wind turbines. |

## I. INTRODUCTION

With the increasing consumption of fossil fuels and the gradual deterioration of environmental problems, there is an urgent need to find a clean and renewable energy source. Wind energy is irreplaceable in energy structures owing to its rapid growth. Wind power accounts for 20% of the world's total electricity, and WT are receiving increasing attention as the core components of wind power generators [1], [2]. Usually, wind power generators are installed in remote areas or offshore areas where traffic is inconvenient, and the

The associate editor coordinating the review of this manuscript and approving it for publication was Yiqi Liu.

gearbox is generally installed in a sky above tens or even hundreds of meters from the ground. In addition, the blades are often subjected to complex alternating impact loads that are transmitted through the main shaft to other critical components of the WT during operation, which makes the daily monitoring and maintenance of the WT difficult. To maximize the economic benefits of the wind farm, the size of the WT is increasing, the capacity of the generator is increasing, and the structure of the corresponding generator has become more complicated. Once a problem occurs, it requires considerable time to check out, which significantly reduces the profits of the wind farm. Therefore, fault diagnosis and maintenance are very important during the operation of WT [3].

The fault diagnosis first needs to obtain the operating data of the WT, and the collected information is usually nonlinear and nonstationary, which contains a lot of noise. In addition, the fault data account for a relatively small amount, and there is a defect in data loss during data transmission [3], [4]. Therefore, the researchers have introduced ML into the fault diagnosis of WT based on this situation. In general, fault diagnosis methods can be divided into four categories [5]:

(a) Method based on a physical model. Building a physical model requires a good understanding of the structure of the WT, but it is almost impossible to achieve a highly nonlinear coupled complex system [5].

(b) Statistical-based approach. In the case of a limited number of samples, the optimization between the learning accuracy of the training samples and the ability to identify arbitrary samples can be optimized to realize the best generalization ability, and utilize the historical data to estimate the changes in WT in the short term [6].

(c) TML-based methods. The known fault samples are represented by the mapping relationship between the input and output of the diagnostic model. In addition, these methods can well characterize the nature of the fault data, which further improves the accuracy of fault diagnosis [7].

(d) A technique based on a hybrid model. Two or more different fault diagnosis models are employed to form a new diagnostic method that can achieve high fault diagnosis accuracy by taking advantage of various diagnostic networks [8], [9].

In the early days, some TML methods have been applied to fault diagnosis, such as Bayesian, decision tree, support vector machine, and random forest, and these methods have some identical diagnostic processes in fault diagnosis. First, the obtained data are preprocessed, the noise in the received signal is reduced, and the abnormal values in the data are addressed. Then, the preprocessed data are sampled or grouped according to different approaches, and the faulty feature extraction is performed on the grouped data. Subsequently, feature selection based on experience and specific diagnostic issues is applied to form a feature vector for fault diagnosis [10]–[12].

However, these methods have some limitations. Usually, the obtained data are nonstationary and contain a large amount of noise. After processing, there is still much noise in the signal, which significantly influences the extraction of fault features. Therefore, more advanced signal processing techniques are required to process the collected data. At the same time, relying on human experience when making feature selection absolutely affects the accuracy of the fault diagnosis, especially for some newly generated faults and insufficient understanding of how the fault develops in the early stage [13]. Furthermore, when the acquired data contain some data that have never appeared or the data distribution has changed, the diagnostic accuracy of the trained model is severely reduced, which causes the diagnostic network to retrain and waste much time [14].

To solve the above problems, researchers have employed artificial neural networks (ANN) for fault diagnoses such as BPNN, RBF, SOM, ART, and ELM. The basic units of the ANN are the neurons and weights between neurons. ANN can complete the learning of the target task in the training process, while simultaneously optimizing its organizational structure to represent the information so as to retain the information in the data to a large extent. When a new fault occurs, it only needs to adjust the weight of the connection between some neurons or increase the number of neurons. Thus, the local adjustment of the trained model can be used for fault diagnosis of new problems. Therefore, retraining of the entire diagnostic model can be avoided and the time consumption is reduced. In addition, the neural network can process data in a parallel manner with a fast calculation speed and high calculation accuracy, which is suitable for building online diagnostic models [15].

Although the diagnostic accuracy of ANN is superior to that of TML methods, there are also some flaws. Most ANNs have only one layer of the hidden layer, which cannot fully exploit the information contained in the data, and some information may be lost during the learning process [16]. In addition, the ability to search for optimum results in the parameter space is limited and cannot provide accurate results [8], and the optimization of existing ANN requires further study.

Based on the basis of ANN, the researchers introduced deep learning (DL), which includes CNN, DBN, SAE and RNN. In practical terms, DL is just a subset of the ML. However, there are two main differences between them. The first difference between them is the way in which the data are presented in the system. ML algorithms almost always require structured data, whereas DL networks rely on layers of neural networks. Another difference is that DL networks do not require human intervention because multilevel layers in neural networks place data in a hierarchy of different concepts, which ultimately learn from their own mistakes.DL uses a greedy learning algorithm to establish the mapping relationship between inputs and outputs through layer-by-layer nonlinear learning, obtain high-dimensional feature representation under different working conditions, and integrate network training and fault diagnosis processes together. For complex nonlinear problems, a more abstract representation of the original data is obtained by increasing the number of layers, and these features have excellent generalization capabilities. In addition, DL can directly learn the acquired data, thereby eliminating the dependence on human experience, and the fault diagnosis model can be transformed into a life prediction model by changing the activation function of the output layer.

However, there is still some potential for improving DL. Because DL diagnoses a fault by multi-layer nonlinear fitting, the computational complexity of the method is increased. Hence, it is necessary to reduce the complexity of the network computing. Most of the current DL methods are like a "black box," which cannot understand the role of each step in the learning process. Therefore, it is necessary to combine the visual learning method to show the results of each stage of DL to improve the method. Another severe problem is that the current DL fault diagnosis method is only applicable to a specific situation or a certain type of situation. When the method is applied to other components or similar problems, the diagnostic accuracy can be significantly reduced.

To solve the problem of reduced diagnostic accuracy in similar situations, researchers have employed transfer learning (TL) to DL, which enables a well-trained model or method to have better generalization capabilities and maintain good diagnostic accuracy in different mechanical systems. However, the negative transfer learning phenomenon generated by the implementation process and how to make full use of the data of the target domain require further research.

It can be seen that, to overcome the different challenges in the field of fault diagnosis, researchers have proposed different methodologies over the past decades. From TML and ANN, to DL and TL, the performance of diagnosis models has been improved with the development of advanced methodologies. Meanwhile, for WT, the fault diagnosis approaches based on ML have rarely been mentioned in the existing works. Therefore, a review is needed to summarize the current research progress in this field. This review makes several contributions to the literature:

(a) The development of WT fault diagnosis method is summarized into four parts, which represent the evolution of ML from TML, ANN, and DL to TL.

(b) A distribution diagram is provided for the discussion of ML methods applied for WT fault diagnosis. The challenges for research on ML methods for WT fault diagnosis are summarized for future research directions.

(c) To the best of our knowledge, this is the first time that such a comprehensive review of fault diagnosis methods and applications specialized in WT is proposed.

The rest of this paper is organized as follows. In Section II, some TML methods are summarized. Section III gives a review of the applications of ANN methods. Section IV mainly focuses on the approaches of DL. Section V introduces TL method and its applications, and Section VI provides a discussion. Conclusions are enclosed in section VII.

## II. TML METHODS APPLIED FOR FAULT DIAGNOSIS OF WT

### *A. SUPPORT VECTOR MACHINE (SVM)*

#### 1) THEORETICAL BASIS

SVM is a pattern-recognition method based on the principle of structural risk minimization. SVM is usually based on a limited sample to find the best generalization ability by finding the optimal compromise between model complexity and learning ability. It attempts to find the maximum margin between the two data categories and then determines the hyperplane that is in middle of the maximum margin.

For a dataset $\{(x_1, y_1), \cdots, (x_l, y_l)\}$ and $y_i \in \{-1, 1\}$, SVM training attempts to find a parameter $w$ and a parameter $b$ of the linear decision function $f(x) = wx + b$ defining the optimal hyperplane. Considering two points $x_1$ and $x_2$ with $f(x_1) = 1$ and $f(x_2) = -1$, the margin equals to:

$$margin = \frac{f(x_1) - f(x_2)}{\|w\|} = \frac{2}{\|w\|}. \tag{1}$$

Thus, maximizing the margin is equivalent to minimizing $\frac{\|w\|}{2}$ or $\frac{\|w\|^2}{2}$. Then, to achieve the optimal hyperplane, the SVM solves the following optimization problem:

$$\begin{aligned} &\min_{w,b} \frac{1}{2} w' w \\ &s.t.\ y_i(w' \cdot x_i + b) \geq 1, \quad \forall i = 1, 2, \cdots, l. \end{aligned} \tag{2}$$

The transformation of this optimization problem into its corresponding dual problem gives the following quadratic problem:

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j), \\ &s.t. \sum_{i=1}^{l} y_i \alpha_i = 0, \quad \alpha \geq 0\ \forall i = 1, 2, \cdots, l, \end{aligned} \tag{3}$$

where $\alpha_i$ is the Lagrange multiplier. The solution of the previous problem gives the parameter $w = \sum_{i=1}^{l} y_i \alpha_i x_i$ of the optimal hyperplane. Hence in dual space, the decision function becomes:

$$f(x) = \sum_{i=1}^{l} \alpha_i y_i (x_i \cdot x) + b. \tag{4}$$

The SVM maps the eigenvectors of the low-dimensional space to a high-dimensional space through appropriate kernel functions, and constructs the optional linear classification hyperplane in the assigned area to classify different types of points. However, the classical SVM is suitable for the two-classification problem, but it is normally a multi-classification problem, and the original SVM needs to be modified to adapt to the multiclass problem by transforming it into a serial of two-class problem as shown

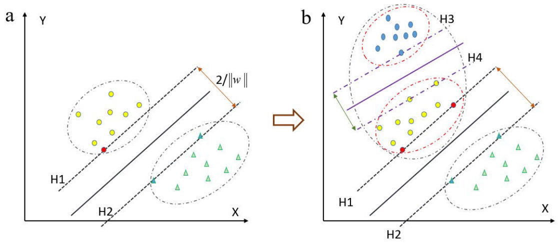


**FIGURE 1.** SVM for classification. (a) SVM for two-classification, (b) SVM for multi-classification.

in Fig 1. Currently, there are primarily four types of SVM for the multi-classification problem: one against all (OAA), one against one (OAO), directed acyclic graph (DAG), and hierarchical SVM (H-SVM).

In the OAA algorithm, a class of samples is distinguished for each classifier, so K of the two-class SVM is needed to identify K class samples. When applying the OAA method, the training time increases proportionally with the increment of the sample, and the sample classification accuracy decreases with the accumulation of the test error rate by multiple classifier algorithms. The OAO algorithm recombines any two of the multiple categories and designs corresponding two-class SVM for each combination. Then, the unknown samples categories are specified according to the amount of the corresponding division of each classifier result. The disadvantage of the algorithm is that the number of classifiers generated increases rapidly with the number of classes, and when there are too many classification categories, the training speed decreases exponentially. The DAG algorithm is a directional graph without a closed loop, which is consistent with OAO during the training phase. For the problem of distinguishing N types of samples, only N-1 steps are required to complete the classification. Compared with the OAO classification, the DAG improves the classification speed. Nevertheless, the algorithm does not study the impact of sample error propagation on subsequent generations. The H-SVM algorithm is similar to DAG. When the morphology of the hierarchy is close to that of the complete binary tree, the method has an ideal training speed and requires fewer classifiers [17].

### 2) APPLICATIONS

Owing to the complexity of the WT system, only some typical faults are discussed in this section.

(a) Fault diagnosis applied to the gearbox of WT.

In Ref. [18], a WT fault diagnosis method was proposed based on a diagonal spectrum and clustering binary tree SVM. Comparing the test results with the general SVM and radial basis function (RBF) neural networks, the method presented in this paper achieved higher diagnostic accuracy with fewer fault samples. The sample attributes were reduced by using a rough neighborhood set in Ref. [19]. The rough neighborhood set ensured that important attributes were added to the reduced attribute set using a forward search method to avoid the loss of essential characteristics. Compared with SVM without rough neighborhood set, the proposed method reduced the computational time and improved the accuracy of fault diagnosis by 6%. In Ref. [6], the optimal penalty factor and kernel function parameters were obtained by cross optimization to construct an SVM classifier with the lowest structural risk and good classification effect. In Ref. [20], Cuckoo search optimization (CSO) algorithm was applied to optimize the kernel function parameters in the SVM. The CSO algorithm is a novel particle swarm optimization (PSO) algorithm, and the diagnostic results of SVM based on the CSO algorithm were increased by 2.5%, 3.5% and 6.5% respectively, as compared with the traditional SVM and SVM based on PSO and KNN. In Ref. [21], the proposed structure selected SVM based on DAG overcame the unreasonable distribution of nodes, which led to poor diagnosis results. By considering the similarity between the nodes, the nodes were selected to maximize the difference between the two categories. However, the proposed method does not guarantee that the chosen results have a minimum diagnostic error rate. In Ref. [22], multi-fault SVM classifiers composed of OAO algorithms can achieve fault diagnosis under multiple operating conditions, which improves the generalization ability of the proposed method. In addition, the current signal is used for fault diagnosis in this method so that the signal acquisition cost is reduced under the premise of improving the signal reliability. According to the characteristics of vibration fault signals in the gearbox of WT, the feature vectors obtained by wavelet decomposition of the gearbox vibration signals in Ref. [23] were input into the multi-class least squares SVM (LSSVM) constructed by the OAO method for model training and fault diagnosis. A novel fault diagnosis method based on manifold learning and Shannon wavelet SVM was presented by Tang *et al.* [24] for WT transmission systems. The successful fault diagnosis application in a WT gearbox proved that the performance of the proposed method is effective.

(b) Fault diagnosis applied to WT bearing.

In Ref. [25], a model of the WT output power was established using the regression SVM. When the WT fails, the generator output power gradually decreases with the gradual increase of the fault, and the residual between the actual value and the predicted value exceeds the standard threshold. Thus, potential failure can be detected through a continuing change in the trend of the residual. Gao *et al.* [26] proposed a novel WT fault diagnosis method based on integral extension load mean decomposition multiscale entropy and LSSVM which was aimed at the nonstationary and nonlinear characteristics of WT vibration signals. The feature vectors were obtained by ensemble empirical mode decomposition (EEMD) [27] and input into the model for training which were optimized by inner cluster distance (ICD). The ICD can measure the degree of separation between different categories in the characteristic space, which can be used to optimize the kernel function parameters in the SVM. The experimental results showed better diagnostic performance than the EMD-ICD-SVM and EEMD-ANN. In Ref. [28], the problem of low-speed rotating bearing fault diagnosis was solved using an alpha stable distribution (ASD). The proposed approach combined the EMD to obtain the characteristic parameters, and chose the most sensible and stable characteristic parameters of the faults that were input into the diagnostic network, and optimized the LSSVM by using PSO. The test results demonstrated that the proposed method can not only achieve fault diagnosis of low-speed rolling bearings' damage position and degree, but also has better diagnostic accuracy and operational efficiency than other methods. In Ref. [29], the variational mode decomposition (VMD) was optimized using the quantum chaotic fruit fly optimization algorithm for the fault diagnosis of the bearing in the WT. The fault features were then grouped into two-dimensional, vectors that were sent to the RVM for fault diagnosis, and the experimental results demonstrated that the proposed method is effective.

(c) Fault diagnosis applied to the rotational imbalance of the WT.

In Ref. [30], a fault diagnosis method for a direct-drive wind turbine based on support vector machine (SVM) and feature selection was presented. Five direct-drive wind turbine experiments were carried out, the features of which were analyzed. Then the sensitive time-domain feature parameters in the horizontal and vertical directions of the vibration signal in the five conditions were selected and used as feature samples. The method was effective in identifying the fault of wind turbine and had good classification ability and robustness to diagnose faults in direct-drive WT. In Ref. [31], the kernel FCM algorithm was used to estimate the degree of similarity between test samples, and the PSO algorithm was employed to optimize the parameters of the kernel function. Finally, the optimized eigenvectors were used to train and test the optimized multiclass fuzzy SVM model constructed using the OAO method. The experimental results demonstrated that the proposed approach is an effective fault diagnosis method. In Ref. [32], the obtained representative features were selected using the principal component analysis (PCA) method and applied to train the proximal SVM (PSVM). Compared with the probability neural network and learning vector quantization, it requires less time and has a higher diagnostic accuracy.

(d) Fault diagnosis applied in the misalignment of WT.

When a misalignment problem occurs, a severe dynamic load is generated between the high-speed shaft of the transmission and the generator shaft, increasing the axial and radial vibrations of the shaft. In addition, the bearing oil leakage, high temperature and loosening of the fastening bolts affect power generation. In Ref. [33], the dual-tree complex wavelet transform (DTCWT) was used to demodulate and reconstruct the signal to obtain the feature vector. Then the PSO algorithm was employed to optimize the kernel function parameters and the penalty factor in the SVM, and construct a multi-class SVM using the OAO method. Nevertheless, this method was only for simulation analysis, and actual cases were not applied to examine the effectiveness of the approach. In Ref. [34], a method of heterogeneous information fusion was proposed to solve the problem of misalignment in WT. First, the feature vectors were obtained by fusing the fault characteristics of the multi-source signals, and dimensionality reduction processing was performed using t-distributed stochastic neighbor embedding (t-SNE). Subsequently, the LSSVM was optimized by using the artificial bee colony algorithm while performing network training. The t-SNE algorithm is based on the conditional probability nonlinear reduction algorithm, which is able to represent high-dimensional information using two-dimensional or three-dimensional data so that it could be used as a visualization algorithm.

#### 3) SUMMARY

SVM maps feature vectors of high-dimensional space by selecting appropriate kernel functions and thus achieving fault classification.

It is particularly suitable for nonlinear processing data and high-dimensional samples. It can obtain higher diagnostic accuracy with fewer and simpler samples, and has good global optimization and generalization capabilities. Nevertheless, it is worth noting that the SVM still has some defects:

(a) The selection of the kernel function and penalty factor is essential for the diagnosis accuracy of SVM.
(b) SVM is suitable for solving the binary problem, for the multi-classification problem and processing a large number of samples, the performance of SVM needs to be improved.
(c) SVM converges to a local minimum easily. Therefore, the SVM should be further optimized and improved.
(d) When a new type of fault occurs, the established diagnostic model should be able to update itself.

### B. DECISION TREE (DT)

#### 1) THEORETICAL BASIS

The DT uses an attribute test of the data to achieve classification. Generally, a DT contains a root node, several intermediate nodes, and several leaf nodes. Each leaf node represents a category corresponding to the decision result. Each intermediate node represents an attribute test. The considering scope of each test is within the bounds of the last decision result, and the contained samples based on the outcome of the attribute test are divided into sub-nodes. The root node contains the complete set of samples, and the path from the root node to each leaf node corresponds to a decision test sequence. A greedy algorithm is applied to the decision tree in the construction process to optimize the current classification effect optimal. There are three forms of termination conditions for decision tree recursion: (a) all samples of the current node belong to one category and do not need to be divided; (b) the current attribute set is empty, or all samples have the same value of all attributes and cannot be separated; (c) the current sample set is empty and cannot be divided.

#### 2) APPLICATIONS

In Ref. [35], the fault diagnosis decision table was obtained from the past fault samples and discretized by the FCM, which was optimized by the rough set theory and the maximum cluster ratio. The optimized decision table was then used together with the C4.5 algorithm to train the DT. The experimental results showed that the proposed method can effectively reduce the computing load of fault diagnosis and improve the diagnostic accuracy. Vamsi *et al.* [36] proposed a method utilizing the multi-source signals to acquire fault features for the gearbox fault diagnosis problem in WT. Then the features were chosen by the C4.5 algorithm, and the dominant features were selected as the feature vector and sent to the SVM for fault diagnosis model training and verification. In Ref. [37], [38], the J48 algorithm was used to extract the fault features, and the features that were most sensitive to the fault were selected as the feature vector, which was sent to fuzzy Q-learning for fault diagnosis in Ref. [37]. The experimental results proved that the performance of the method was better than that of the ANN and SVM. In Ref. [38], the features were sent to a classifier based on the best-first tree algorithm and functional trees algorithm, and it was proved that the fault diagnosis rate of the method in WT blade reached 91.67% in 10-fold cross validation. In Ref. [39], the DT method was used for fault diagnosis, and a graphical relationship between the fault source and the fault was established. Then, the reasoning between the fault and the fault source was learned based on the existing fault samples, and the learning results were utilized for fault diagnosis. Both the diagnostic accuracy and the confirmation time of the fault were improved.

#### 3) SUMMARY

The DT can deepen the understanding of faults in WT based on the learning of existing fault samples, and does not require much processing of the data when establishing the DT. In addition, it can deal with the data containing noise, and is not sensitive to information loss. Nevertheless, it is easy to converge to a local minimum, and the risk of overfitting the data is high, and cannot be used to establish an online diagnostic network. Moreover, when there are too many categories need to be classified, the diagnostic accuracy may decrease.

### C. BAYES

#### 1) THEORETICAL BASIS

Bayesian decision theory is a method for implementing decisions based on prior knowledge of the conditions. The fundamental principle of the Bayes classifier is the Bayes rule, as follows:

$$P(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{\sum_{j=1}^{c} p(x|\omega_j)P(\omega_j)}. \tag{5}$$

The Bayes rule indicates how the information of known probability density functions, $p(x|\omega_i)$, and a priori probabilities, $P(\omega_i)$, can be used to calculate the posteriori probability, $P(\omega_i|x)$. The minimum-error- rate classification can be achieved by use of the Bayes discriminant functions as follows:

$$g_i(x) = \ln P(x|\omega_i) + \ln P(\omega_i), \quad where\ i = 1, 2, \cdots, c. \tag{6}$$

This expression above can be readily evaluated if the densities $p(x|\omega_i)$ are normal distribution, that is, if there is a distribution $p(x|\omega_i) \sim N(\mu_i, \sum_i)$. So in this case, we have:

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \sum_i (x-\mu_i) - \frac{d}{2}\ln 2\pi \\ -\frac{1}{2}\ln|\sum_i| + \ln P(\omega_i), \tag{7}$$

where $\mu_i$ is the mean and $d \ln 2\pi$ is a constant. All the necessary information of each class and feature cluster is contained in the mean vector and covariance matrix. The center of each cluster is determined by the mean vector and the shape of the cluster using the covariance matrix. As for sorting tasks, in cases where all correlation probabilities are known, Bayesian decision-making considers how to choose the best class label based on these probabilities and misjudged losses. Because each sample has the possibility of being mislabeled, the expected loss due to a misclassification of the sample which is the conditional risk of an example can be calculated based on the posterior probability. To minimize the risk of overall misjudgment, it is necessary to reduce the conditional risk of each sample, which is to select the category label that minimizes the conditional likelihood for each sample based on the Bayesian decision rule. The posterior probability used in the Bayesian decision rule, which is immeasurable in actual operating conditions is usually obtained indirectly by Bayes' theorem. Nowadays, the naïve Bayes classifier is widely used to solve the problem of joint probability between attributes when calculating the posterior probability. The naïve Bayes classifier uses the "attribute conditional independence assumption" which means that for all known categories, it is assumed that all attributes are independent of each other. The semi-naïve Bayes classifiers are an improvement of the naive Bayesian classifiers by providing proper consideration of the interdependence of attributes. [7]

#### 2) APPLICATIONS

(a) Fault diagnosis applied to the gearbox of WT.

Song *et al.* [40] conducted a comparison study on three types of Bayesian diagnostic models constructed based on SCADA, which include the bin method, the multivariate normal distribution-based method, and the Copula method. Examining the same data with the three diagnostic models showed that the Copula method can provide more diagnostic information. Moreover, the Bayesian diagnostic model was compared with the traditional energy curve-based diagnostic method to illustrate the superiority of the Bayesian diagnostic network. Li *et al.* [41] proposed a method based on tunable

Q-factor wavelet transform-morphological component analysis (TQWT-MCA) and a sparse Bayesian iteration algorithm combined with a step-impulse dictionary to address the issue of high-speed remote transmission and large-capacity data storage. The proposed method solves the problem of data loss and confusion during data transmission by considering the interaction of data between multiple channels. Yu *et al.* [42] tried to establish a decision table using the characteristic relation and then used an assignment reduction algorithm to remove redundant features, and the remaining features were sent to a flexible naïve Bayesian classifier for fault diagnosis. The experimental results prove that the proposed method can achieve a planetary gearbox fault diagnosis with incomplete diagnostic information, reduce computational complexity, and enhance reasoning accuracy. Zhong *et al.* [43] combined correlation analysis with the Hilbert-Huang transform (HHT) to extract eigenvectors, and then eigenvectors were sent to a pairwise-coupled sparse Bayesian extreme learning machine for model training and fault diagnosis. Comparing the method with pairwise-coupled probabilistic neural networks and pairwise-coupled RVM, the results demonstrated the effectiveness of the proposed method.

(b) Fault diagnosis applied to the bearing of WT.

Herp *et al.* [44] developed a method that assumes that the fault features extracted from the SCADA data obey the Gaussian distribution in the feature space, and the fault diagnosis model was established by learning the fault samples. The experimental outcomes demonstrated that the method can detect faults on average 33 days in advance, but the method has low fault diagnosis accuracy and relies heavily on high-quality samples. D. Wang [45] proposed an approach to combine the Infogram with the novel Bayesian inference to improve the wavelet filtering to determine the optimal wavelet parameters and apply them for fault diagnosis. Two instance studies proved that the proposed Bayesian inference method is convergent and provides more fault signatures than the Infogram. Li *et al.* [46] utilized the PSO algorithm to optimize the importance of different signals in the process of extracting fault features. The PCA algorithm was then used to select the most sensitive fault features that were subsequently sent to the three-tier Bayesian belief network for fault diagnosis. The effectiveness of the proposed method was verified through experiments. Yu *et al.* [47] presented a fault feature extraction method based on mean multi-granulation decision-theoretic rough set (MMG-DTRS) and non-naive Bayesian classifier (NNBC). To begin with, fault features were obtained by the MMG-DTRS, and the representative feature dimension was then reduced by the attribute reduction algorithm, and finally the fault diagnosis was performed by the NNBC. In addition, the choice of optimized bandwidth in NNBC was used to ignore the assumption of attribute independence, and the joint probability density function was applied to replace the edge probability density function to make the diagnostic model more realistic.

#### 3) SUMMARY

Bayesian networks diagnose faults by minimizing the risk of separation. The diagnostic model can make the fault diagnosis fast and straightforward. For multi-classification problems, the computational complexity does not increase significantly. In addition, in the case of independent distribution of attributes, the diagnosis effect is particularly useful, and it is not sensitive to data loss. Nevertheless, the network cannot work when there is an error category that has never been recorded in the data, and the diagnostic accuracy decreases when the input data attributes are not independent of each other. Moreover, the diagnostic system needs to know the prior probability in advance, which is usually subjected to a hypothetical model. If the model selected is unreasonable, the diagnostic accuracy decreases.

### D. HIDDEN MARKOV MODEL (HMM)

#### 1) THEORETICAL BASIS

HMM is based on the Markov chain. The state at the next moment is only related to the current state and has nothing to do with the previous state. There are two variables in HMM. The first one is a state variable, which indicates the state of the system at a precise moment because the system often changes between these states, so the state variables of the system are usually discrete spaces with some possible values and usually assume that the state variables of the system are not measurable. The second state variable is the observed variable, which is the observed value of the system at a particular time. The observed values may be discrete or continuous. The transition between states in the basic HMM depends on a certain probability, which is the state transition probability matrix, and assumes that the probability matrix and the state in the diagnostic model do not change over time.

An $N$ states' HMM can be expressed by:

$$\lambda = (N, M, \pi, A, B), \tag{8}$$

where $N$ is the number of state of HMM and the state $b_t \in \{S_1, S_2, \cdots, S_N\}$ at moment $t$. $M$ is the number of possible observations for each state. $\pi$ denotes to probability of the original state and $\pi = (\pi_1, \pi_2, \cdots, \pi_N)$. Note that $\pi_i$ can be described as:

$$\pi_i = P(q_i = S_i), \quad 1 \leq i \leq N, \tag{9}$$

and $\pi_i$ satisfies to normalization condition, i.e. $\sum_{i=1}^{N} \pi_i = 1$. $A$ is state transition probability matrix,

$$A = (a_{ij})_{N\times N}, \quad a_{ij} = P(q_{t+1} = S_j | q_t = S_t), \\ 1 \leq i, j \leq N, \quad \sum_{j=1}^{N} a_{ij} = 1. \tag{10}$$

$B$ is observed probability matrix and

$$B = (b_{jk})_{N\times M}, \quad b_{jk} = P(O_t = V_k | q_t = S_j), \\ 1 \leq j \leq N, \quad 1 \leq k \leq M, \sum_{k=1}^{N} b_{jk} = 1. \tag{11}$$

HMM model can also be simplely remarked as $\lambda = (\pi, A, B)$. Moreover, in a classification problem, the diagnostic model is usually estimated based on the observed sequence, and maximizes the probability of occurrence of the observed sequence [48], [49].

#### 2) APPLICATIONS

(a) Fault diagnosis applied to the gearbox of WT.

Li *et al.* [50] presented a new optimal Bayesian control approach to predict the early fault of a partially observable gear shaft system subjected to deterioration and random failure. The optimal maintenance policy represented by a multivariate Bayesian control scheme based on a hidden semi-Markov model (HSMM) was developed. The method determines whether there is a fault by detecting the posterior probability, and the model can be applied to online diagnosis once the boundary of the posterior probability is determined by learning. Long-term vibration data were collected from a 3MW WT based on two-year observations [51]. The observation thresholds of various signals and the correlation coefficients between the signals were determined and committed to the HMM for fault diagnosis. The experimental results indicated that the success rate of the method was as high as 95%. However, considering the continuous monitoring of the acquired data, the threshold in the method needs to be adjusted in real time, and when a new state occurs, the state transition matrix should update in time.

(b) Fault diagnosis applied to the bearing of WT.

In order to tackle the issue that HMM is not very efficient in accuracy and sensitivity of fault diagnosis [52], the fuzzy scalar quantization method was used to reduce the influence of outliers in the data which makes the HMM more sensitive to fault characteristics. Gao *et al.* [53] proposed a method to obtain the feature vector by combining the local mean decomposition with the mutual information method and the false nearest neighbor, and then the features were used for HMM training and diagnosis. The experimental results show that the proposed method can effectively identify the different faults of the rolling bearing. Liu *et al.* [54] introduced a hybrid generalized HMM-based condition monitoring approach for rolling bearings where interval valued features were used to efficiently recognize and classify machine states in the machine process. The PCA technique was applied to reduce the dimensionality of features that were obtained by VMD, and the remaining fault features were sent to the generalized HMM for fault diagnosis. The experimental results proved the superiority of the method in signal processing and fault diagnosis.

#### 3) SUMMARY

The HMM is a graphical diagnostic model based on the Markov chain. This indicates the reasoning process for fault diagnosis. However, because of the nature of the Markov chain, it is only related to the state of the previous moment. Hence, the historical data cannot be fully utilized, and the probability of the state transition matrix remains unchanged in the diagnostic model, which is not in line with the real situation. Moreover, the topology in the diagnostic model is sometimes ambiguous, and it is challenging to create a precise topology.

### *E. RANDOM FOREST (RF)*

#### 1) THEORETICAL BASIS

RF is a typical representative of ensemble learning, completing learning tasks by building multiple learners. The RF makes multiple learners parallel based on the decision tree to learn one problem at the same time. The results of these learners decide on the final results. Moreover, in the process of learning in an RF, the current partitioning attribute is obtained by randomly selecting k attributes in the attribute set, and then the optimal attributes among the k attributes are chosen. In this way, not only the diversity of samples but also the diversity of attributes improve the generalization ability of RF algorithms.

#### 2) APPLICATIONS

(a) Fault diagnosis applied to the gearbox of WT.

Gan and Jiao [55] proposed a fault diagnosis method for WT' gearbox based on an improved genetic algorithm RF classifier. The acquired feature vectors were filtered by the PCA algorithm to form feature vectors, which were sent to the RF for fault diagnosis. In addition, the number of RF learners in the method and the number of attributes in each node partition attribute set were optimized using the genetic algorithm (GA). The experimental results indicated that the method is more effective than SVM and traditional RF. A gearbox fault diagnosis method based on deep RF fusion of acoustic and vibration signals was introduced by Li *et al.* [56]. The obtained feature vectors were sent to the deep Boltzmann machine for high-dimensional fault feature learning, and the learned fault features were fused by RF and used for fault diagnosis. The experimental results indicated that the diagnostic accuracy of the method was as high as 97.68% for the 11 different condition patterns.

(b) Fault diagnosis applied to the bearing of WT.

Han and Jiang [57] utilized VMD to obtain a feature vector that was sent to the RF for fault diagnosis. The experimental results showed that the approach achieved higher accuracy than SVM, GA-SVM, and PSO-SVM with less time. Qin [58] combined ensemble empirical mode decomposition (EEMD) and RF to achieve a fault diagnosis. The feature vectors obtained by EEMD are subsequently sent to the RF for fault diagnosis. Five cross-validations proved the effectiveness of this method. To address the issue of traditional feature extraction methods, it is difficult to accurately extract fault information, and there is a serious problem of information redundancy in fault diagnosis. Jia *et al.* [59] used the two-dimensional signal correlation information combined with the complex EMD to obtain the fault features, and then used the Gini index to measure the importance of each feature to perform feature screening, and finally sent the features into the RF for fault diagnosis. Simulation and experimental results shows that the method can effectively extract fault features.

#### 3) SUMMARY

RF is a fault diagnosis model based on the decision tree. Because it learns the same problem through multiple learners, it can obtain better diagnostic results than the general learner. Through different sampling methods and selecting different partitioning attributes, the robustness of RF is enhanced, which makes the method more generalizable. Moreover, the RF is not sensitive to outliers in the data. However, because it contains multiple learners, the training process is complicated. And it requires much more calculation time. If there is intense noise in the processed data, the method may suffer from the risk of overfitting. In addition, when selecting a learner, there is no clear guideline to specify which learner is better at diagnosing a problem.

### *F. CLUSTERING ALGORITHM (CA)*

#### 1) THEORETICAL BASIS

The CA is an unsupervised learning algorithm that attempts to divide the unmarked samples into several disjoint subsets, each of the subsets integrates into a "cluster" as shown in Fig. 2. The classification goal is that the samples within the cluster should be as similar as possible, and the differences between the clusters are expected to be as substantial as possible. The most widely used CA is the k-means and FCM algorithms.

The K-means algorithm uses the squared loss function to measure the similarity of the samples in the cluster and selects the cluster center by minimizing the squared loss function. The loss function can be defined as the sum of squares of errors between each sample and the center point of the cluster to which it belongs:

$$J(c,\mu) = \sum_{i=1}^{M} \|x_i - \mu_{c_i}\|^2, \tag{12}$$

where $x_i$ represents the $i$th sample, $c_i$ is the cluster to which $x_i$ belongs, $\mu_{c_i}$ represents the center point corresponding to the cluster, and $M$ is the total number of samples.

The FCM algorithm determines the label of the sample by calculating the membership of each sample and the cluster center. By optimizing the objective function, it obtains the membership degree of each sample point to all class centers, so as to automatically classify the samples. If the computed values exceed a certain threshold, a new sample marker is added to accommodate the change of the sample. Suppose we have a dataset $X$, and we want to classify the data in it. If these data are divided into $c$ classes, the corresponding $c$ class centers are $c_i$, and each sample $x_j$ belongs to a certain class, and the membership of $c_i$ is determined as $u_{ij}$. Then we define a FCM objective function and its constraints as follows:

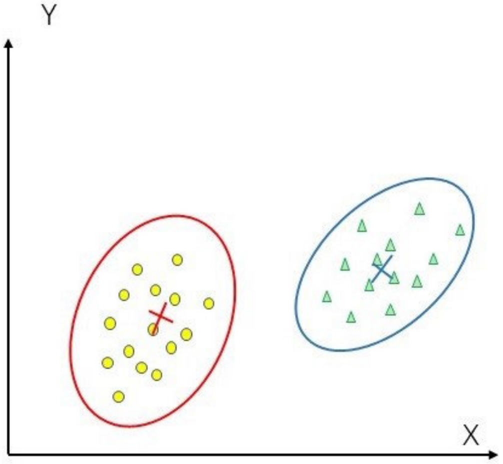


**FIGURE 2.** **Clustering algorithm schematic diagram.**

$$J = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m} \|x_j - c_i\|^2, \tag{13}$$

$$\sum_{i=1}^{c} u_{ij} = 1, \quad j = 1, 2, \cdots, n, \tag{14}$$

where $m$ is a factor of membership, generally values to 2, and $\|x_j - c_i\|$ represents the Euclidean distance from $x_j$ to the center point $c_i$.

#### 2) APPLICATIONS

(a) Fault diagnosis applied to the gearbox of WT.

A feature extraction method based on multi-fractal approximate entropy and subtractive clustering has been proposed [60]. First, the multi-fractal spectrum was combined with the approximate entropy to obtain the feature vector, and then the subtractive clustering was used to process the feature vector to make the initial clustering center, and FCM was applied for fault diagnosis. Among them, subtractive clustering can effectively solve the problem in which FCM easily falls into the local optimum. And it improves the convergence speed. Wu *et al.* [61] proposed a method to use kernel FCM to improve the clustering result of fuzzy c-means clustering for nonlinear data. Additionally, a gravitational search algorithm was proposed to solve the randomness of clustering centers. The results showed that concurrent faults could be effectively diagnosed.

(b) Fault diagnosis applied to the bearing of WT.

The K-means CA was used to process the outliers of the data [62], and then the auto-associative neural network was combined with the residual approach to approximate the data distribution to a normal distribution, and Hotelling multivariate quality control charts were used for fault identification. However, the method of processing data needs to be improved, and the diagnostic model cannot be updated in time. Because the accuracy of fault diagnosis is limited considering a single operating condition, Zhang *et al.* [63] combined the historical state of the SCADA data with the K-means CA to divide the health state into four subspaces. In each operational state subspace, according to the Gaussian mixture model, a basic model of health status and a health index was established based on the Mahalanobis distance to assess the health status of the online WT. However, this method does not adequately consider changes in the operating conditions under regular operation. A novel cluster-contraction stage-wise orthogonal matching-pursuit approach for bearing fault information extraction was presented [64]. The clustering contraction mechanism was added to the Stage-wise Orthogonal-Matching-Pursuit (StOMP) algorithm and the selected atoms were filtered twice during atomic search, making the condition number of the support set more reasonable, thus realizing amelioration of the pathological equation in weight determination.

#### 3) SUMMARY

The CA achieves classification by calculating the different membership degrees of each sample in the cluster center. This method can manage massive quantities of data with high efficiency. However, the number of categories needs to be given in advance, and the selection of the initial cluster center has a significant influence on the training of the model. With the new data input, the cluster center of each class should continually adapt, which increases the amount of computation of the method. In addition, it may converge to a local minimum when processing big data.

### *G. SUMMARY OF TML*

From the above introduction, it can be found that the application of TML methods for fault diagnosis of WT is pervasive. In fact, TML-based fault diagnosis methods are a big domain of many technologies. Many valuable methods that have been applied to fault diagnosis in addition to those we have summarized. For instance, [65] developed a surrogate model method, namely the modified Kriging-based moving extremum framework (MKMEF) to efficiently perform probabilistic analyses of the structural dynamic response. To improve the dynamic reliability analysis of complex structures such as turbine blisk, [66] proposed a moving extremum surrogate modeling strategy (MESMS) method. In [67], two different modified multi-extremum response surface basis models (MRSM) were proposed for dynamic nonlinear responses of failure capacities for turbine blisk responses. Usually, the information on the fault samples is analyzed to integrate the representative fault features based on past fault samples.  A summary of the applications of TML for WT fault diagnosis is presented in Table 1.

**TABLE 1.** **The summary of applications of TML to WT fault diagnosis.**

| Methodologies | Monitoring Components | References |
|---|---|---|
| SVM | Gearbox | Li et al. [17], Liu et al. [18], Ning et al. [19], Agasthian et al. [20], Lihui et al. [21], Cheng et al. [22], Liu et al. [23], Tang et al. [24] |
| | Bearing | Li et al. [25], Gao et al. [26], Zhang et al. [27], Qing et al. [28], Xu et al. [29] |
| | Rotation imbalance | An et al. [30], Hang et al. [31], Malik et al. [32] |
| | Misalignment | Xiao et al. [33], Xiao et al. [34] |
| DT | Gears, Generation system | Wang et al. [35], Vamsi et al. [36], Malik et al. [37], Joshuva et al. [38], Liu et al. [39] |
| Bayes | Gearbox | Song et al. [40], Li et al. [41], Yu et al. [42], Zhong et al. [43] |
| | Bearing | Herp et al. [44], Wang et al. [45], Ke et al. [46], Yu et al. [47] |
| HMM | Gearbox | Márquez et al. [48], Man et al. [49], Li et al. [50], Sung et al. [51] |
| | Bearing | Liu et al. [52], Gao et al. [53], Liu et al. [54] |
| RF | Gearbox | Gan et al. [55], Li et al. [56] |
| | Bearing | Han et al. [57], Qin et al. [58], Jia et al. [59] |
| CA | Gearbox | Li et al. [60], Wu et al. [61] |
| | Bearing | Yang et al. [62], Zhang et al. [63], Song et al. [64] |

To reduce the complexity of fault diagnosis in high-dimensional space and the time spent in model training, the acquired fault features are subjected to a dimensionality reduction process to remove redundant information and obtain more sensitive fault features. Then, a peculiar type of WT fault is trained in the model based on these fault characteristics. Nevertheless, there are some issues with this type of method:

(a) Because the acquired data are usually nonlinear and non-stationary, the data contain considerable noise interference. Extracting the fault features of these data while eliminating noise requires advanced signal processing methods. Nevertheless, the existing method can deal with these problems, simply because the approach is too complicated and time consuming, so it is not suitable for online fault diagnosis.

(b) Fault features usually obtained in the time, frequency, and time-frequency domains. Although different fault features provided by different domains in the same algorithm can contain more fault information at the same time, it increases the complexity of the algorithm, and the selection of fault features in a specific fault usually depends on prior knowledge. Therefore, it is necessary to further study the domain to select for feature extraction and the fault features in the domain.

(c) The usually trained model only works for the particular problem merely, when the data structure changes or new issues arise, the model needs to be retrained, which means that it needs more time, and the accuracy of the diagnosis result decreases. Hence, it was not possible to achieve real-time updates of the model.

(d) Environmental information, climate change, and other factors should be considered when selecting fault feature vectors to ensure the comprehensiveness of the information obtained.

## III. ARTIFICIAL NEURAL NETWORKS (ANN)

The ANN is a data-driven fault diagnosis method based on a biological model. ANN establishes a system of adaptive neurons with a broad scope of parallel interconnections, and fault diagnosis is achieved by nonlinear fitting of the input samples and output results. A typical representative of an ANN is a multilayer perceptron as shown in Fig 3. It consists of an input layer, an output layer, and a hidden layer in the center. Because of their high self-learning capability, ANN-based models could automatically learn better diagnosis knowledge with fewer prior knowledge than TML. Therefore, they can be regarded as more advanced technologies than TML-based methods.

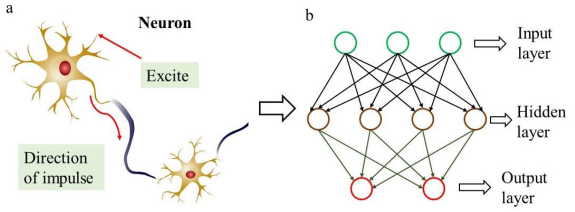


**FIGURE 3. A typical representative of an ANN. (a) Human neuron information transmission schematic diagram, (b) Typical ANN schematic diagram.**

Fig 3 (b) depicts a fully connected neural network, where each neuron is connected to each neuron in the adjacent layer. The input layer neurons only accept external inputs. The hidden layer neurons not only accept the information from the previous layer but also generate an output to the next layer by comparing the total received inputs with the threshold of the neuron. In addition, the effect produced by the hidden layer was invisible. The output layer neurons produce outputs based on the inputs obtained through an activation function. This section introduces several widely used ANNs.

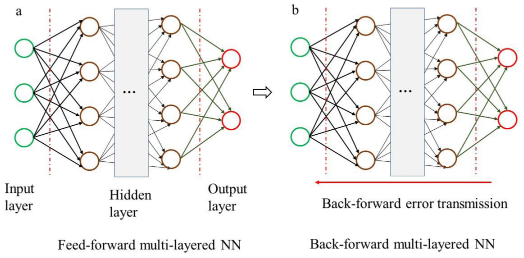

a
Input layer
Hidden layer
Output layer
Feed-forward multi-layered NN
b
Back-forward error transmission
Back-forward multi-layered NN

**FIGURE 4. Backpropagation learning rule. (a) Feed-forward multi-layered ANN, (b) Back-forward multi-layered ANN.**

### A. BACK PROPAGATION NEURAL NETWORK (BPNN)

#### 1) THEORETICAL BASIS

The BPNN is a multilayer feedforward network trained by an error back-propagation algorithm. The training process includes two parts: the forward propagation of the diagnosis network and the reverse fine adjustment of the network parameters. The error propagation in the opposite direction is to distribute the output error through the hidden layer to the input layer, giving out the error to each layer unit. The learning rule applies the gradient descent method to continuously adjust the weights and thresholds of the network through backpropagation as shown in Fig 4, thus minimizing the sum of squared errors of the network. Given the training dataset $\{x_i, y_i\}_{i=1}^{m}$ with $m$ samples, where $x_i \in \mathbb{R}^d$ includes $d$ features and $y_i \in \mathbb{R}^l$ includes $l$ health states, the output of the $h$th hidden layer is expressed as:

$$(x_i^h)_j = \sigma^h(\sum_{i=1}^{h-1} \omega_j^h \cdot x_i^{h-1} + b_j^h), \quad j = 1, 2, \cdots, n_h, h = 1, 2, \cdots, H, \qquad (15)$$

where $(x_i^h)_j$ is the output of the $j$th neuron in the $h$th hidden layer, and $x_i^0 = x_i$, $n_h$ is the number of neurons in the $h$th hidden layer, $\sigma^h$ represents the activation function of the $h$th hidden layer, $n_{h-1}$ is the number of neurons in the $(h-1)$th hidden layer, $\omega_j^h$ is the weights between the neurons in the previous layer and the $j$th neuron in the $h$th hidden layer, and $b_j^h$ is the bias of the $h$th hidden layer. The predicted output of BPNN is:

$$(\hat{y})_k = \sigma^{out}(\sum_{i=1}^{n_H} \omega_j^{out} \cdot x_i^H + b_j^{out}), \quad k = 1, 2, \cdots, l, \qquad (16)$$

where $(\hat{y})_k$ is the predicted output of the $k$th neuron in the output layer, $\sigma^{out}$ is the activation function of the output layer, $\omega_j^{out}$ and $b_j^{out}$ are respectively the weights and bias of the output layer. When given a certain training sample$x_i, y_i$, the optimization objective of BPNN aims to minimize the error between the predicted output and the target one by:

$$\min_{\omega,b} E_i = \frac{1}{2}\sum_{k=1}^{l}[(y_i)_k - (\hat{y}_i)_k]^2. \qquad (17)$$

#### 2) APPLICATIONS

To address the problems of highly nonlinear and nonstationary signals of WT, a fault diagnosis method for direct-drive WT was proposed by An *et al.* [68] used a BPNN. The time-domain feature parameters of the vibration signals in the horizontal and vertical directions were considered in the method. Test samples were utilized to verify the validity of the BPNN model, and showed a higher diagnostic accuracy. In Ref. [69], the Levenberg-Marquardt (L-M) algorithm was utilized for the BPNN, and the faults of the gearbox, wind pressure difference system and generator in the WT were diagnosed. The MATLAB simulation experiment was used to determine the fault characteristics that were most sensitive in differentiating faults, and it was indicated that it is feasible to use neural networks for the fault predictive diagnosis of WT. Wang *et al.* [70] proposed a method based on a BPNN to address the WT converter circuit fault diagnosis online. The converter circuit model of the doubly fed WT has strong nonlinearity, which makes it difficult to diagnose faults online. Data acquisition and data normalization were performed using critical points in the voltage or current signals that might be faulty. Then, the data with fault information were sent to a four-layer BPNN for fault diagnosis. The field results proved the effectiveness of the proposed method. Han *et al.* [71] proposed a tabu search method to optimize the BPNN for predicting the output energy of the WT. Owing to the ability of global optimization, the tabu search method can improve the BPNN deficiency that can quickly converge to the local optimal value. By inputting appropriate parameters, the method can improve the convergence speed while improving the diagnostic accuracy.

### 3) SUMMARY

BPNN has been widely applied in fault diagnosis because of its maturity. Under the assumption that there are enough neurons in the hidden layer, the three-layer neural network can approximate any function with arbitrary precision. Network training sustains self-learning and self-adjustment capability. Moreover, this method has fault tolerance ability. When an individual neuron has an error, the impact on the overall result is not significant, and there is a specific generalization ability for the same type of problem. However, the convergence speed of BPNN is slow, and it is easy to converge to the local optimal value, and there is a certain risk of overfitting. The selection of the initial weight in the method and the selection of the learning rate in the fine-tuning process need to be further studied.

## B. EXTREME LEARNING MACHINE (ELM)

### 1) THEORETICAL BASIS

The ELM is an improvement of BPNN, also a feedforward neural network, so it has an input layer, a hidden layer, and an output layer. The connection weights of the input layer and the hidden layer and the threshold of the hidden layer are set randomly or artificially and do not need to be updated during the learning process. The weighted connections of the hidden and output layers are determined by solving the system of equations and cannot be updated once determined.

If the training dataset is given as

$$\{x_i, t_i | x_i \in \mathbb{R}^D, t_i \in \mathbb{R}^m, i = 1, 2, \cdots, N\}, \tag{18}$$

where $x_i$ is $i$th data instance, $t_i$ is the label of $i$th data instance, and such set denotes to all training data, the output of hidden layer can be written as,

$$H(x) = [h_1(x), h_2(x), \cdots, h_L(x)] \tag{19}$$

where $h_i(x)(i = 1, 2, \cdots, L)$ is output of the ith hidden layer node, which is not unique. Generally, $h_i(x)$ can be described as follows:

$$h_i(x) = g(w_i, b_i, x) = g(w_i x + b), \quad w_i \in \mathbb{R}^D,\ b_i \in \mathbb{R} \tag{20}$$

where $g(w_i, b_i, x)$ is the activation function and Sigmoid function and Gaussian function are commonly used. $w_i$ and $b_i$ are parameters of hidden layer nodes. Then the output of the "generalized" single hidden layer feedforward neural network ELM is:

$$f_L(x) = \sum_{i=1}^{L} \beta_i h_i(x) = H(x)\beta \tag{21}$$

where $\beta = [\beta_1, \beta_2, \cdots, \beta_L]^T$ is the output weights between hidden and output layers

### 2) APPLICATIONS

(a) Fault diagnosis applied to the gearbox of WT.

Li *et al.* [72] proposed a method that combines VMD with the kernel ELM (KELM) for rolling bearing fault diagnosis. The feature vectors obtained by the VMD method were sent to the KELM for fault diagnosis. Simultaneously, the PSO algorithm was employed to optimize the penalty factor and kernel function parameters in the KELM. The experimental results indicated that the proposed method performed better than BPNN, SVM, and ELM. An intelligent WT gearbox diagnosis approach using VMDEA (optimize VMD parameter with differential evolution algorithm (DEA)) and ELM was reported by Isham *et al.* [73]. The mode number and the balancing parameter in the VMD were optimized based on DEA [72], and the signal was demodulated by the optimized VMD to obtain the feature vector for the ELM training. The experimental results indicated that the diagnostic accuracy of the method was improved by 10% for the ELM-based diagnostic method, and the diagnostic accuracy of the method was improved by 5%-10% for the VMD-based diagnostic method. A one-dimensional feature vector obtained by HHT was sent to a pairwise-coupled sparse Bayesian ELM for model training [43]. Experiments demonstrated that this method can consume a relatively short time to obtain higher diagnostic accuracy. To address the challenges of the dynamical and high-dimensional data generated from the WT generator system, Yang *et al.* [74] proposed a new fault diagnosis scheme composed of multiple ELMs. A multiple ELM in a hierarchical structure was utilized to achieve feature extraction and dimensionality reduction of data, and the last layer of the ELM was set as a fault classifier. Compared with other fault diagnosis methods that use a combination of wavelet packet transform (WPT), time-domain statistical feature (TDSF) and kernel principal component analysis (KPCA), the diagnostic accuracy of the proposed method was improved by 5%-10%. In Ref. [75], cloud computing technology was adopted for the condition monitoring of WT, and the compressed sensing method was used for data transmission, which can reduce the data transmission amount while ensuring the data security. The obtained feature vector was sent to the hierarchical ELM for fault recognition, and experiments proved the robustness of the proposed method.

(b) Fault diagnosis applied to the transmission chain of WT.

To solve the weakness (weak generalization ability, low diagnostic rate) of traditional fault diagnosis with feedforward neural networks, a fault diagnosis method based on an improved extreme learning machine (IELM) was proposed [76]. The sample set was updated by the new fault sample and the historical sample similarity, which guarantees that the diagnostic network can update the network according to the change in the fault sample while reducing the calculation time. Compared with SVM, ELM, and fixed-size sequential ELM, the proposed method has better diagnostic accuracy. Wang *et al.* [77] proposed a novel dual-ELM-based fault diagnostic framework for feature extraction and fault pattern recognition. The acquired fault characteristics are sent to the dual-ELM for fault diagnosis. The dual-ELM contains only two basic ELMs, one for calculating the number of faults and the other for indicating the types of faults that may be involved. In addition, the network structure of the method is relatively small, and the ELM operation speed can be effectively utilized. In Ref. [78], an online sequential ELM method was proposed for fault diagnosis. The obtained fault features were normalized using a physical kinetic energy correction model to eliminate the effects of different speeds. The short-term and long-term fault diagnosis of WT through SCADA data proved the effectiveness of the method, and the method can update the model in real time according to the data change.

#### 3) SUMMARY

ELM is normally deemed as an improvement of BPNN because the learning speed of the algorithm is greatly improved as compared with the BPNN. Moreover, the algorithm performs complex learning by increasing the number of hidden layer neurons. Usually, the number of hidden layer neurons is connected to the categories that need to be distinguished so that they can adapt to new situations. At the same time, this method has better generalization ability. However, this method has only one layer of hidden neurons, and thus, the learning ability is limited.

### C. RADIAL BASIS FUNCTION NEURAL NETWORK (RBFNN)

#### 1) THEORETICAL BASIS

The RBFNN is a single hidden layer feedforward neural network that uses a radial basis function as the activation function of the hidden layer neuron, which is a local response function. The radial basis function(RBF) is a real-valued function whose value depends only on the distance from the origin or arbitrary point $c$, and point $c$ is called the central point. That is, the RBF can be described as follows:

$$\begin{aligned}\phi(x) &= \phi(\|x\|),\\ or\ \phi(x,c) &= \phi(\|x-c\|).\end{aligned} \tag{22}$$

Gaussian function $\phi(x) = \exp(-r^2/2\sigma^2)$, multiquadric function $\phi(x) = \sqrt{x^2+c^2}$, and inverse Multiquadric function $\phi(x) = 1/\sqrt{x^2+c^2}$ are some tipical RBFs.

In the application of this method, the number of center points and the position of the center points should be determined first, which are the number and location of neurons in the hidden layer. In addition, the method calculates the distance between the input sample and the hidden layer point as the input of the radial basis function, and the output layer is the linear combination of the output of the hidden layer neuron function. Moreover, the weights and biases in the neural network were determined from the process of learning the samples.

#### 2) APPLICATIONS

In Ref. [79], for the fault diagnosis of blades of WT, the fault features were first extracted based on historical data, and then the features were selected using the J48 algorithm. Subsequently, the acquired features were employed to train the sequential minimal optimization (SMO) algorithm and simple logistic algorithm (SLA), BPNN, logistic algorithm (LA), and RBFNN. The experimental results indicated that the SLA had most significant effect, and the correct rate eached 93.67%. A data-driven fault diagnosis and isolation method was proposed for fault diagnosis of actuators in WT [80]. The obtained feature vectors were sent to the BPNN, RBFNN, DT, and KNN for parallel training, and then the results of the respective learners were fused to obtain the results. Simulation results and Monte Carlo sensitivity analysis have proved the effectiveness of the proposed method.

#### 3) SUMMARY

The RBFNN has strong nonlinear fitting ability, which can deal with the law that is difficult to analyze in the system. Moreover, it has fast convergence speed and better generalization ability. Moreover, its performance for classification problems is better than BPNN. However, the quality of the learning data and the selection of samples significantly influenced on the learning effect. There is a problem of information loss in the learning process, and the reasoning process of the method cannot be explained.

### D. SELF-ORGANIZING MAP (SOM)

#### 1) THEORETICAL BASIS

SOM is a competitive learning model belonging to unsupervised neural network, which maps high-dimensional inputs to low-dimensional space while maintaining the topological structure of input data in the high-dimensional area, in other words, similar sample points in high-dimensional space are mapped to the adjacent neurons in the network output layer. The output layer neurons of SOM are arranged in a matrix in a two-dimensional space, and each neuron has a weight vector. After accepting the input vector, the network determines the winning neurons in the output layer and the position of the input vectors in the low-dimensional space. The vectors of the winning neurons and their neighboring neurons were then continuously adjusted to reduce the distances between these weight vectors and the current input samples. The purpose of

SOM training is to find an appropriate weight vector for each neuron to maintain the topology.

The SOM process involves for major components, initialization, competition, cooperation and adaptation. In the initialization process, all connection weights were initialized with small random values.

In the competitive process, if the input space is $D$ dimension, the input data can be written as $x = \{x_i : i = 1, \ldots, D\}$ and the connection weights between the input units $i$ and the neurons $j$ in the computation layer can be written as $w_j = \{w_{ji} : j = 1, \ldots, N; i = 1, \ldots, D\}$ where $N$ is the total number of neurons. The discriminant function can be defined as the squared Euclidean distance between the input vector $x$ and the weight vector $w_j$ for each neuron $j$,

$$d_j(x) = \sum_{i=1}^{D}(x_i - w_{ji})^2. \tag{23}$$

The neuron whose weight vector comes closest ro the input vector is declared the winner.

In the cooperative process, a similar topological neighbourhood for the neurons in SOM is defined as,

$$T_{j,I(x)} = \exp(-\frac{S^2_{j,I(x)}}{2\sigma^2}), \tag{24}$$

where $S_{ij}$ is the lateral distance between neurons $i$ and $j$ on the grid of neurons, $I(x)$ is the index of the winning neuron. A special feature of the SOM is that the size $\sigma$ of the neighbourhood needs to decrease with time which is defined as $\sigma(t) = \sigma_0 \exp(-t/\tau_\sigma)$ where $\sigma_0$ and $\tau_\sigma$ are super parameters.

In the adaptive process, the winning neuron and its neighbours have their weights updated. In practice, the appropriate weight update euation is

$$\Delta w_{ji} = \eta(t)\ \cdot T_{j,I(x)} \cdot (x_i - w_{ji}), \tag{25}$$

where t is a time epoch and $\eta(t) = \eta_0 \exp(-t/\tau_\eta)$ is the learning rate. $\eta_0$ and $\tau_\eta$ are super parameters.

#### 2) APPLICATIONS

To extract fault features of WT from complicated nonlinear time-varying system, an novel algorithm that combines a modified 1ocal discriminant basis(LDB) algorithm and SOM was presented [81]. The obtained signal passed through the LDB algorithm, and the received characteristic information was sent to the SOM network to transform the one-dimensional feature into two-dimensional(2D) feature, thereby improving the signal separation degree. Subsequently, the 2D elements were sent to the BPNN for fault diagnosis. In Ref. [82], the fault characteristics of the converter output voltage or current were obtained by wavelet transform, and then the fault features were sent to the SOM network for network training and fault diagnosis. The results showed that the performance of the proposed method was effective. Yang *et al.* [83] proposed a hybrid SOM–PCA method for the fault diagnosis of bearings in WT. The fault features were first extracted, and then PCA was used to select the appropriate fault features for the SOM network training. The results indicated the efficiency of themethod in the field of unsupervised classification of most faults. Kramer *et al.* [84] presented a method from neural computation that can serve as forecasting and monitoring techniques for WT energy prediction and supervision problem. The support vector regression method was proposed to predict the energy output, and the SOM method was used to reduce the dimensionality of the high-dimensional data. The experimental parts were based on real wind energy time series data from the National Renewable Energy Laboratory (NREL) western wind resource dataset, and experimental case studies were described to validate the proposed method, which showed the efficiency of fault detection and diagnostic in experimental data such as the unsupervised classification of most faults. Wang *et al.* [85] proposed a new abnormality detection and prediction technique based on heterogeneous signals and information. The output power signals and wind turbin downtime event information were collected from the SCADA system, and then transported to a linear mixture SOM classification for network training and fault diagnosis. Experiments showed that this method is better than the traditional SOM method. Gil *et al.* [86] presented indicators of non-expected behavior in components of a WT for the fault diagnosis of the generator and gearbox. Several diagnostic models were established under normal working conditions, and then the state of WT was judged by whether the actual behavior conformed to the expected behavior. To identify the health status of WT, Blanco *et al.* [87] presented a strategy based on SOM and interpretation-oriented post-processing tools. There was a problem with the same type of fault for various components in WT. Thus, SOM combined with CA was utilized to put the failures of similar behaviors into the same subset. On this basis, a post-processing tool was added to further enhance the degree of aggregation. The experimental results demonstrates the effectiveness of the proposed method.

#### 3) SUMMARY

The output results of the SOM network are easy to understand. This method can also be used for the visualization of an algorithm, and its implementation structure is relatively simple. However, the calculation complexity of this method is relatively high, and it cannot learn a dataset with missing data. When the number of learning samples is small, the order of the input sample has a significant influence on the learning results. Moreover, the setting of the initial state and the selection of parameters influence the convergence speed of this method.

### *E. ADAPTIVE RESONANCE THEORY (ART)*

#### 1) THEORETICAL BASIS

ART is a competitive learning network. The network has a comparison layer, identification layer, identified threshold, and reset module. The comparison layer receives the input sample and transmits it to the recognition layer. Each neuron of the recognition layer corresponds to a pattern category, and the number of neurons can be dynamically increased during the training process to add a new pattern class. Upon having the input signal of the comparison layer, each neuron in the recognition layer calculates the distance between the representative vector of all storage classes and the input vectors, and the smallest distance is the winning neuron. If the similarity between the input vector and the winning neuron is greater than the recognition threshold, the current input sample is marked as the category to which the vector belongs, and the network weight is updated. Subsequently, when accepting similar input samples, the model calculates a higher degree of similarity, which gives the neuron a higher chance of winning. When the similarity between the input vector and the winning neuron is below the threshold, the reset module adds a new neuron in the recognition layer, and its representative vector is the current input vector.

The structure of ART includes layer $F1$ and $F2$, layer $F1$ is the comparison layer and layer $F2$ is the recognition. The neurons of each sublayer of F1 works in a shunt model and operates in an instantaneous balance. $V_i(i = 1, \ldots, M)$ represents the activity of neurons and can be calculated as,

$$\epsilon \frac{d}{dt} V_i = -AV_i + (1-BV_i)J_i^+ - (C + DV_i)J_i^-, \qquad (26)$$

where $A$ and $D$ are constants, $B, C = 0$, and $J_i^+$ represents the stimulation and $J_i^-$ represents the inhibition. The solution of $V_i$ is,

$$V_i = \frac{J_i^+}{A + DJ_i^-}, \qquad (27)$$

The winning neuron $y$ is obtained by layer $F2$ selects the from the result of $F1$, which is calculated as,

$$T_j = \sum_i p_i z_{ji} \qquad (28)$$

$$y_i = \begin{cases} d, & T_j = \max_k\{T_k\}, \ \forall k \text{ not marked}, \\ 0, & otherwise. \end{cases} \qquad (29)$$

where $T_j$ is the similarity of neuron $j$ of $F2$, $z_{ji}$ represents the connection weight between neuron $i$ of $F1$ and neuron $j$ of $F2$, $p_i$ is a parameter of layer $F1$ and $d$ is a super parameter that can be selected according to different systems.

### 2) APPLICATIONS

Yang *et al.* [88] constructed a vibration condition monitoring system for WT bearings based on noise suppression using multi-point data fusion. The feature vectors acquired by the EMD correlation analysis were sent to ART for fault diagnosis. As an improved version of the ART, ART-2 can learn autonomously about emerging issues, and no prior knowledge is required. Through an analysis of the actual and simulated fault vibration signals of wind turbine bearings, the proposed EMD correlation model supplemented with ART-2 data fusion could not only effectively remove white noise and short-term disturbance noise but also extract early weak fault feature frequencies. In Ref. [89], the feature vectors obtained by the discrete wavelet transform (DWT) were addressed by ART-2 for fault diagnosis. The proposed method was applied to the vibration signals collected from a gearbox to diagnose gear-crack faults. The results showed that the relative wavelet energy could effectively extract the signal feature and that the ART-2 neural network could recognize the changing trend from the normal state to a crack fault before the occurrence of a broken tooth fault. Ben *et al.* [90] proposed an online automatic diagnosis approach for WT bearings progressive degradations based on ART-2, which was simulated by the Randall model and combined with the obtained feature vector into ART-2 for fault diagnosis. The use of real measured data from a wind turbine drivetrain proved that the proposed data-driven approach is suitable for online condition monitoring of WT bearings even under real experimental conditions and achieves a better generalization capability as compared to previous works even with noisy measurements. Lee *et al.* [91] presented a model-based fault diagnosis method for detecting and isolating faults in a robot arm control system. When a change in the system occurred, the errors between the system output and the estimated output crossed a predetermined threshold, and once a fault in the system was detected, the estimated parameters were transferred to the fault classifier by ART2 neural networks with uneven vigilance parameters for fault isolation. For the problem of gear fault diagnosis, the soft competitive learning fuzzy ART method was adopted by Wan *et al.* [92]. The Yu norm was introduced to solve the influence of the input sample order in the soft competitive learning fuzzy ART, and the lateral inhibition theory was also applied to solve the problem of modal node chaos. The obtained fault characteristics were sent to the soft competitive learning fuzzy ART for training, and then the weighted voting method was employed to make a judgment based on the correlation selection partial diagnosis results. Experiments demonstrated that the proposed method had better diagnostic accuracy and generalization ability.

### 3) SUMMARY

ART is an incremental learning method that can learn new problems while retaining previous learning experiences. It is an unsupervised learning method that can learn the problem without any prior knowledge and can stably and quickly identify the object that has been determined. In addition, the method can avoid the disadvantages of other algorithms that easily converge to a local minimum. However, the ART needs to set parameters, and it is very difficult to determine the optimal combination of parameters. In addition, ART cannot handle imbalanced datasets and the order of the test data affects the final clustering results.

## F. SUMMARY FOR ANN

ANN makes fault diagnosis faster and has a higher diagnostic accuracy than TML methods. Utilizing the nonlinear fitting of the ANN between the inputs and outputs can better learn

**TABLE 2. The summary of applications of ANN to WT fault diagnosis.**

| Methodologies | Monitoring Components | References |
|---|---|---|
| BPNN | Gearbox, Transmission chain, generator fault | An et al. [68], Ju et al. [69], Zhe et al. [70], Shuang et al. [71] |
| ELM | Gearbox | Li et al. [72],Isham et al. [73], Yang et al. [74], Qian et al. [75] |
|  | Transmission chain | Wu et al. [76], Wang et al. [77], Qian et al. [78] |
| RBFNN | Blades, Actuators | Joshuva et al. [79], Pashazadeh et al. [80] |
| SOM | Bearings, Gearbox | Zhemin et al. [81], Yang et al. [82], Fadda et al. [83],Kramer et al. [84], Wang et al. [85], Angle et al. [86], Alejandro et al. [87] |
| ART | Bearings, Gearbox | Yang et al. [88], Li et al. [89], Ali et al. [90], Lee et al. [91], Wan et al. [92] |

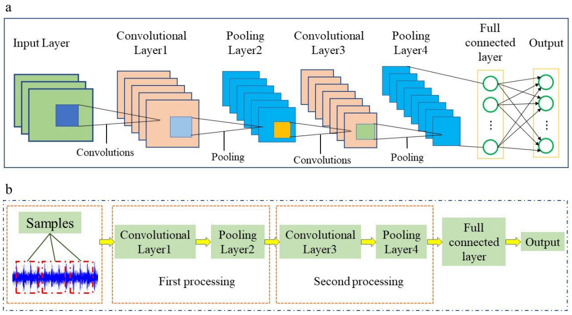


**FIGURE 5. A typical CNN schematic diagram and its application for fault diagnosis.**

the faulty information. When the structure of the input information changes or a new type of fault occurs, the ANN can learn new information structures or problems by modifying the link weights between some neurons, updating the thresholds, or adding new neurons. The ability of ANN to learn indivisible linear problems, especially high-dimensional data, is better than that of TML algorithms. Although the degree of dependence of ANN on prior knowledge is not as high as that of conventional TML algorithms, it still needs to rely on manual experience in feature extraction and feature selection. Meanwhile, other than the typical ANN-based methods reviewed in this section, there are many ANN-based works for fault diagnosis that are in the minority but have shown good performance. For example, [93] proposed an extremum response surface method by introducing generalized regression neural network (GRNN) and a multi-population genetic algorithm (MPGA).This is an effective work that provides a learning-based reliability analysis method for complex equipment.

The summary of applications of ANN for WT fault diagnosis can be shown in Table 2.

However, the hidden layer in the ANN usually has one layer, which causes the neural network to have limitations in learning data and may miss some vital information in the learning process to affect the accuracy of fault diagnosis. Another problem is that there is no precise standard for determining the number of neurons in each layer of the network. If the number of neurons is too small, the learning effect is impacted, and the accuracy of the fault diagnosis is subsequently reduced. However, if the number of neurons is too large, it takes much more time is required to train the model, and the probability of overfitting increases, which reduces the generalization ability of the proposed method.

## IV. DEEP LEARNING (DL)

To solve the problem that the acquisition process of fault features still relies on the manual experience in the ANN as well as only the superficial level features are learned and the extraction of deep-level features is insufficient, researchers introduced DL fault diagnosis for WT. The DL directly learns the acquired data through a multidimensional neural network to obtain high-dimensional features, which are a more essential representation of the data structure. Therefore, the generalizability of the diagnostic method is improved. DL combines two processes of feature acquisition and fault diagnosis, which reduces the dependence on advanced signal processing technology and diagnostic experience to improve the accuracy of fault diagnosis. In addition, this method has no special requirements for data and can process high-dimensional and nonlinear data that cannot be addressed in the past. The most commonly used DL networks are convolutional neural networks (CNN), deep belief networks (DBN), stacked auto-encoders (SAE), and recurrent neural networks (RNN).

### A. CONVOLUTIONAL NEURAL NETWORKS (CNN)

#### 1) THEORETICAL BASIS

A CNN is a typical feedforward neural network. It essentially aims to build multiple filters that can extract the features of the input data. Through layer-by-layer convolution and pooling of these filters, the topological features hidden in the data are extracted step by step, and finally, the characteristics of the input data are obtained during the process of translation, rotation, and scaling. The CNN is usually composed of an input layer, convolution layer, pooling layer, fully connected layer, and an output layer, and the convolution layer and the pooling layer are alternately arranged as shown in Fig. 5 (a). Fig. 5(b) shows the normal procedure for fault diagnosis with CNN. The signal samples are first processed by the stacked convolutional layer and pooling layer to obtain local features, and then the fully connected layer is used to transform local features into global features, and finally the probability distribution of the sample is output The convolution layer of the CNN includes many feature planes, each of which contains many neurons, and each neuron connects to a local area of the upper feature layer through the convolution kernel, which is a weight matrix. In addition, CNN obtains fault features through convolution operations and share weights in the same feature plane, which reduces the complexity of the model and makes the model easier to train. The pooled layer follows the convolutional layer and has a structure similar to that of the previous convolutional layer, which aims to obtain spatially invariant features by reducing the resolution of the feature surface. Moreover, pooling operations reduce the number of neurons, thereby reducing the computational complexity of the network model [94].

#### 2) APPLICATIONS

(a) Fault diagnosis applied to the gearbox of WT.

Two-dimensional feature vectors were obtained by different methods and sent to the CNN for fault diagnosis [95]. The proposed method was verified and compared with other methods to demonstrate that the proposed method has a robust ability to suppress noise while processing data and can obtain a diagnostic accuracy of 99.3%. An algorithm was proposed by Monteiro *et al.* [96] to reduce the training time by 80% without affecting the accuracy of the CNN diagnosis. The obtained features were sent to the CNN for training, and the output results were sent to the SVM for fault recognition. The algorithm was improved by explaining the results of the DL in the decision-making stage. To solve the problem that the CNN learning process cannot be explained, the layer-wise correlation propagation method was proposed to make the CNN learning process visible [97], which demodulates the two-dimensional signal output after CNN learning into a series of pixel points and quantizes them to obtain the feature variables that have the greatest impact on fault diagnosis.

(b) Fault diagnosis applied to the bearing of WT.

In Ref. [98]–[100], CNN was compared with several traditional ML methods, both the simulation and experimental results indicated that CNN is much better in terms of diagnosis accuracy and network convergence speed. However, this method still has the possibility of misclassification, so more sensors were needed to obtain data to reduce the probability of errors. In the process of fault diagnosis, PSO algorithm was used to optimize the number of convolution kernels and learning rate of CNN [101]. Comparing the trained model with ANN network and SVM, it was proved that the method can achieve higher diagnosis accuracy, but the calculation efficiency of the method needs to be further improved. Guo *et al.* [102] proposed an intelligent fault diagnosis method for bearings with variable rotating speed based on Pythagorean spatial pyramid pooling (PSPP) and CNN. Two dimensional fault features were obtained by CWT and PSPP, and sent to CNN for fault diagnosis. The accuracy of PSPP and CNN can reach 99.11%, which is much better than other fault diagnosis methods. Moreover, the model trained at a certain speed can be applied to fault diagnosis under all operating conditions, but if more environmental information is considered, the accuracy of fault diagnosis needs to be further improved.

Chen *et al.* [103] presented a wind power generation fault diagnosis approach based on a DL model using the Internet of Things (IoT) with clusters. The representative features of these data can be obtained through the collection and fusion of a variety of data. When these representative features were used in the training of DL, the trained network had higher diagnostic accuracy and smaller prediction error, which was demonstrated by the experimental results. In Ref. [104], the acquired signal was input into the CNN for network training. By learning the one-dimensional signal directly, the loss of data can be avoided, and the fault diagnosis accuracy of the network was improved as well. Zhuang *et al.* [105] introduced a novel fusion diagnosis method for rotor system faults based on DL and multi-sourced heterogeneous monitoring data. A multi-source CNN (M-CNN) was used to learn one-dimensional or two-dimensional signals. The high-dimensional features obtained were subsequently fused by t-SNE and the fused features were sent to M-CNN for fault diagnosis. The experimental results proved the feasibility of using multi-source heterogeneous data for fault diagnosis.

#### 3) SUMMARY

Compared with the traditional ANN, CNN reduces the variable parameters in the network training process through the weight sharing network in the convolutional layer, which reduces the complexity of the network model and avoids overfitting of data, thus improving the generalization ability of the CNN. At the same time, the pooling operation used in the CNN structure dramatically reduces the number of neurons in the model, and the translation invariance of the input data makes the CNN more robust. However, this method has several drawbacks. A large amount of data are required in the network training process, which increases the computational complexity of the method. During the pooling process, some critical information may be lost, and the method is generally incomprehensible for the data learning process. In addition, the CNN can only process input data with a fixed length, so it is necessary to further exploit the potential advantages of this method in combination with other theories.

### B. DEEP BELIEF NETWORKS (DBN)

#### 1) THEORETICAL BASIS

DBN is composed of a several restricted Boltzmann machines (RBM). An RBM contains only one visible unit and one hidden unit. The visible and hidden layers were bidirectionally connected. The visible layer units $v = \{v_1, v_2, \cdots, v_m\}$ and the hidden layer units $h = \{h_1, h_2, \cdots, h_n\}$ are bidirectionally connected. As an energy-based model, the variables $v$ and $h$ are subject to the joint configuration as follows:

$$E(v, h, \theta) = -\sum_{i=1}^{m}\sum_{j=1}^{n}\omega_{ij}v_i h_j - \sum_{i=1}^{m} b_i v_i - \sum_{j=1}^{n} a_j h_j, \quad (30)$$

where $\theta = \{\omega, a, b\}$ represents the parameters of RBM. Then, the marginal distribution of the visual units can be calculated as:

$$P(v|\theta) = \frac{1}{Z(\theta)} \sum_{h} \exp[-E(v, h, \theta)], \tag{31}$$

where $Z(\theta) = \sum_{v,h} \exp[-E(v, h, \theta)]$ is the partition function. There is no connection between neurons in the same layer, and there is a weight between any connected neurons to indicate the connection strength. The multi-layer perceptron of multiple nonlinear operation hidden layers is used to represent the input data in a distributed way, which can learn the essential features of the dataset in the case of limited samples, and achieve high-level feature representation and extraction of data. In classification learning, supervised learning is usually used, and its core is to use a greedy algorithm to optimize the connection weight in DL layer by layer. In other words, unsupervised learning is used to train layer-by-layer to effectively mine the fault features in the signal to be diagnosed to ensure as much feature information as possible when the feature vector maps to different feature spaces. Then through the BPNN of the last layer, the weights in the trained DBN are adjusted in reverse through supervised learning. Generally, all layers of the DBN are considered as a whole by using the method of random gradient descent to minimize the network training error and optimize the fault identification ability of DBN [106], [107].

### 2) APPLICATIONS

Yu *et al.* [108] proposed a radically data-driven fault detection and diagnosis (FDD) method based on DBN for WT. In a wind turbine benchmark Simulink model, DBN was compared with the existing four model-based algorithms and four data-driven algorithms, and the superiority of DBN for the diagnosis of WT faults was proved. In Ref. [109], the obtained fault features were sent to the DBN for fault diagnosis. The proposed method was compared with SVM, SOM, BPNN and Mahalanobis distance (MD) to diagnose the same fault, and the experimental results proved that DBN is superior to other methods. In addition, the paper pointed out that the performance of the method could be further tested using mixed marked and unmarked data. Liu *et al.* [110] proposed a fault diagnosis method for WT gearbox based on DBN and vibration signals. Before the vibration signals were sent to the DBN network for learning, the data were processed by the batch normalization method, which can effectively reduce the probability of overfitting and improve the convergence speed of the method. Through experiments, the improved method performed better than ordinary DBN and BPNN. An approach using an optimized DBN for rolling bearing fault diagnosis was presented by Shao *et al.* [111]. The obtained vibration data were directly input into the DBN network for learning. At the same time, the PSO algorithm was used to determine the optimal number of neurons in each hidden layer, learning rate, and momentum of the DBN. Through a comparison with SVM and ANN based on simulation and experimental data, the accuracy and robustness of the optimized DBN were proved. Li *et al.* [56] proposed a combination method to achieve a gearbox fault diagnosis. After the vibration signals of gearbox were processed, they were sent to the DBN for DL to obtain high-dimensional fault features, which were fused to the RF for fault diagnosis. The accuracy of the method was 97.68% through 11 types of tests under different working conditions. To solve the problem of gradient disappearance in reverse fine tuning, improved sigmoid units have been proposed [112]. By combining the traditional sigmoid units with the merits of unsaturation from leaky rectified linear units, when the absolute value of the gradient was greater than a certain threshold value, the continuous gradient was replaced by a constant gradient to avoid the problem of gradient disappearance. To further improve the diagnosis accuracy, the optimized Morlet wavelet transform was also used to process the signal, which was sent to the DBN for learning and training. After testing, the fault diagnosis accuracy of this method reached 96.32%.

### 3) SUMMARY

DBN is a probability generation model. The method can eliminate the dependence on a large amount of signal processing technologies and experience to complete the adaptive extraction of fault features, and fault diagnosis of WT. In addition, the method has no periodic requirements for data and has strong versatility and adaptability. Moreover, the method can process high-dimensional, nonlinear data, which can effectively avoid dimensional disasters or insufficient diagnostic capabilities. However, this method has some shortcomings in this method. This method is usually only capable of processing one-dimensional data or one-dimensional feature signals, but the learning effect of two-dimensional signals or feature vectors declines, which wastes a lot of time in the forward training process of greedy learning, and it needs to choose the appropriate method to reduce the calculation time.

## *C. STACKED AUTOMATIC ENCODER (SAE)*

### 1) THEORETICAL BASIS

SAE is similar to DBN, which is used to replace RBM in DBN with AE (autoencoder). AE is a three-layer unsupervised neural network that includes an input layer, hidden layer and output layer. Generally, the number of neurons in the input and output layers is the same, while the number of neurons in the hidden layer is less than that in the input layer and output layers to obtain the characteristic representation of the input data. AE is used to achieve the maximum reduction of data through two processes, encoding and decoding, while minimizing the data reduction error.

Given the dataset $\{x_i, y_i\}_{i=1}^{m}$ with $m$ samples, the represented features $h_i$ are defined as:

$$h_i = f_\theta(x_i) = \sigma_f(\omega^T \cdot x_i + b), \tag{32}$$

where $\sigma_f$ is the activation function of the encoder network, and $\theta = \{\omega, b\}$ is the training parameters of the encoder network. The reconstructed sample $\hat{x}_i$ can be obtained by the decoder network, which is expressed as follows:

$$\hat{x}_i = g_{\theta'}(h_i) = \sigma_g(\omega'^T \cdot h_i + b'), \tag{33}$$

where $\sigma_g$ is the activation function of the decoder network, and $\theta' = \{\omega', b'\}$ represents the training parameters of the decoder network. In order to reconstruct the original input as well as possible, the optimization objective of AE focuses on minimizing the error between the input samples and the reconstructed ones by:

$$\min_{\theta,\theta'} L(x_i, \hat{x}_i) = \frac{1}{2m} \sum_{i=1}^{m} \|x_i - \hat{x}_i\|^2. \tag{34}$$

The gradient descent algorithm is still utilized for the network training of the AE. In the pre-training stage, with the aim of minimizing the input and output errors, each AE is trained one by one until all the AE are trained, and then the weights and offsets between each layer of AE are adjusted through reverse fine-tuning to extract high-dimensional features from low-dimensional data. However, in the process of reverse fine-tuning, there are problems of gradient disappearance and gradient diffusion [107]. The DAE is a variation of the AE and is an improvement that can deal with the gradient diffusion problem.

#### 2) APPLICATIONS

##### *a: FAULT DIAGNOSIS APPLIED TO THE GEARBOX OF WT*

Chen *et al.* [113] introduced a fault diagnosis method based on rotor current for doubly fed induction generators WT drivetrain gearboxes using frequency analysis and a deep classifier. The processed data were sent to the SAE to extract the deep representative features, and then the multiclass SVM was used for fault diagnosis. Experiments demonstrated that the proposed method achieves a more considerable improvement than traditional diagnostic methods. In Ref. [114], to solve the problem of few gearbox failure samples, a method based on generative adversarial networks (GANs) was proposed. Based on the existing fault samples, a new fault sample with similar data distribution was generated by using the GAN algorithm, and used in SDAE training together with the previous fault samples. By comparing the trained fault diagnosis model with SDAE, SAE, BPNN, SVM and ELM, the validity of the proposed method was proved. Yu *et al.* [115] presented a method based on selective deep SDAE with negative correlation learning (NCL) for gearbox fault diagnosis. The acquired signal was used for forward learning by SDAE followed by reverse fine tuning with NCL, and then the fault diagnosis was completed by the PSO-optimized integrated learning algorithm. By comparing SDAE-NCL with DBN, BPN, SVM, KNN, and RF, the validity of this method was proved. The acquired signal is usually accompanied by a large amount of noise, and the common method can only learn from the data of a specific noise level. To overcome this problem, the AE was trained by the noise level to form stacked multilevel-denoising autoencoders (SMLDAE) [116]. Through experimental analysis, it can be seen that SMLDAE has advantages over MLP, SAE, and SDA, but the paper pointed out that the method of increasing multi-level noise needs to be improved, and the method cannot deal with unbalanced data.

##### *b: FAULT DIAGNOSIS APPLIED TO THE BEARING OF WT*

Lu *et al.* [117] proposed a fault diagnosis method for rotary machinery components using an SDA-based health state identification method. The obtained data were sent to the SDA for the fault diagnosis. By comparing with SVM and RF, it can be seen that the proposed method has many advantages, but at the same time, a large number of samples are needed to further verify the limitations of this method and optimize the DL structure. To improve the accuracy of fault diagnosis and recognition, an intelligent fault diagnosis approach based on sparse deep neural networks was introduced [118]. The data received from the four different states were learned by SAE, and the established diagnosis network was verified by experiments, and the results showed that the diagnostic accuracy of the proposed method was 98%. Shao *et al.* [119] proposed a method to improve the fusion of the depth features. The obtained signals were sent to the DAE and the output results were sent to the CAE (contracting auto encoder) to further extract fault features. Then the extracted fault features were fused in the LPP (locality preserving project) to improve the feature learning ability, and the fused features were sent to softmax for fault diagnosis. Compared with BP and SVM, this method was proven to be effective, and exhibited better performance than the standard CNN.

#### 3) SUMMARY

SAE is a discriminant generation model. Compared with DBN, this method lacks the strict requirements of layer parameterization, and can achieve high-performance fault diagnosis results through a small amount of sample data and an appropriate classification basis. SDAE can overcome the gradient diffusion problem in the process of backward fine tuning, and it is easier to understand the internal learning process than DBN. However, there are some limitations to this method. Because the number of encoders used should be determined according to different working conditions, when the number of encoders is too large, the training time of the model increases, along with an increase in the risk of overfitting, and there is no clear rule to measure the selected results.

### *D. RECURRENT NEURAL NETWORKS (RNN)*

#### 1) THEORETICAL BASIS

Fig. 6 shows a typical RNN schematic diagram. RNN has both internal feedback and feedforward connections between processing units. The internal feedback connection can maintain the state of the hidden nodes and provide memory for the network. The output of the network is not only related to the current input, but also to the internal state of the previous network, reflecting good dynamic characteristics.

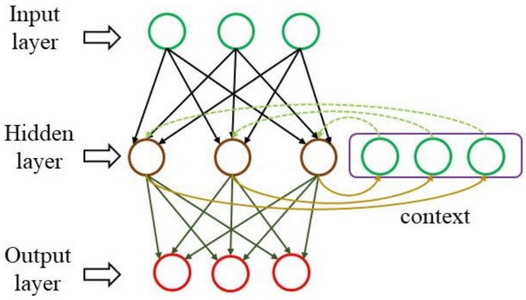


**FIGURE 6.** Typical RNN schematic diagram.

The output $O_t$ and the value of hidden layer $S_t$ can be calculated as:

$$O_t = g(V \cdot S_t), \tag{35}$$

$$S_t = f(U \cdot X_t + W \cdot S_{t-1}), \tag{36}$$

where $t$ is the moment, $U$, $V$, $W$ are parameters of the system network, $X_t$ represents the instance vector. In the whole training process, the same parameter $W$ is used at each time. The RNN fully considers the association between samples, which is reflected by the connections between neural networks. Generally, the neurons of RNN have the same weight and offset, and the RBM or AE can be utilized for pre-training to initialize the network parameters. Then, the output error of each sample is calculated, and the network parameters with the accumulated error are trained [120]. However, in the process of each feedback, some information is lost. When the time accumulated to a certain extent, the initial information degenerated and the gradient vanishing effect appeared.

### 2) APPLICATIONS

As an improvement on RNN, the most widely used RNN is the LSTM neural network, which solves the problem of gradient disappearance in the reverse fine-tuning process of RNN. Because some faults occur slowly and usually require an extended period, conventional methods are challenging to deal with such problems, but LSTM can take advantage of the hidden long-term dependencies in the data. The LSTM network structure is similar to a typical neural network structure and can be divided into an input layer, a hidden layer, and an output layer, which correspond to input gates, hidden gates, and output gates in the network. By controlling the gate, the degree of disturbance of the saved information with the new incoming information can be controlled, and the network assigns corresponding weights and offsets to the currently saved value and the new input value.

To detect faults that occurred at the appropriate time, Talebi *et al.* [121] presented a robust fault detection system (FDS) of wind energy conversion systems (WECS) based on dynamic neural networks. By utilizing a comprehensive dynamic model that contains both the mechanical and electrical components of the WECS, an FDS was suggested by applying a dynamic RNN. The proposed FDS detects faults of different sensors and pitch actuators. By employing an adaptive threshold, FDS robustness was achieved. Because the inherent defect of RNN may cause the phenomenon of gradient disappearance, the LSTM network was used to overcome this defect [122]–[124].The acquired signals were processed and then sent to the LSTM for training, and the fault diagnosis was performed by making full use of the spatial and temporal dependencies in the signals. In Ref. [124], it was pointed out that CNN could be used as a data preprocessing process to enhance the fault diagnosis capability of the LSTM method. In Ref. [125], the SCADA data were divided into several intervals according to the variance, and then the corresponding LSTM networks were trained for each interval. Subsequently, the trained network was used to diagnose the main components of WT, and then the state of WT was evaluated. Although the advantage of this method is that it can evaluate the state of WT, it takes much more time to build multiple models, and it needs to be optimized by other algorithms. In Ref. [126], a residual signal was generated by learning the original signal through LSTM, and then the residual signal was sent to the RF for fault diagnosis. Compared with the other four model-based algorithms and four data-driven algorithms, this algorithm was proven to be effective. However, the proposed algorithm requires a large amount of fault data when it is implemented and cannot update the model automatically.

### 3) SUMMARY

The hidden neurons in the RNN network not only receive the output of the upper layer, but also receive feedback from the neurons of this layer, and can process data of infinite length. Owing to the different structures of the feedforward neural network, it is possible to diagnose slow-developing faults by hiding long-term dependencies in the data. Moreover, it is challenging to deal with the problem of gradient disappearance in the recursive process, but LSTM as an improvement of RNN can solve the problem of gradient disappearance. However, the trade-offs between new input information and current state information require further study, and there are no clear rules on how to choose the optimal number of hidden neurons and the appropriate RNN structure. If the choice is not appropriate, the network may be unstable, turbulent, or chaotic.

## *E. SUMMARY OF DL*

From the above analysis, it can be seen that the DL can directly obtain high-dimensional and in-depth representative features from the original data, which solves the issues of the TML and ANN methods caused by fault feature selection which depends on the manual experience. The summary of applications of DL to WT fault diagnosis is shown in Table 3. Moreover, the high-quality diagnosis results in the cited literature have shown that the deep features learned by deep networks contain more diagnostic information than the TML and ANN, and the DL-based models gain better generalization ability. In addition, the DL-based methods are totally

**TABLE 3. The summary of applications of DL to WT fault diagnosis.**

| Methodologies | Monitoring Components | References |
|---|---|---|
| CNN | Gearbox | Zhou et al. [94], Chen et al. [95], Monteiro et al. [96], John et al. [97] |
| | Bearing | Min et al. [98], Dobie et al. [99], Verstockt et al. [100], Wang et al. [101], Guo et al. [102], Chen et al. [103], Zhu et al. [104], Zhuang et al. [105] |
| DBN | Gearbox | Rui et al. [106], Helbing et al. [107], Yu et al. [108], Tamilselvan et al. [109], Liu et al. [110], Shao et al. [111], Qin et al. [112] |
| SAE | Gearbox | Cheng et al. [113], Wang et al. [114], Yu et al. [115], Jiang et al. [116] |
| | Bearing | Chen et al. [117], Fei et al. [118], Shao et al. [119] |
| RNN | Gearbox | Tang et al. [120], Talebi et al. [121], Qian et al. [122], Yang et al. [123], Lei et al. [124], Sun et al. [125], Li et al. [126] |

data-driven, which means those models could automatically recognize all different fault types if the training dataset is sufficient.

At the same time, DL has some flaws:

(a) There is no uniform rule for the selection of the number of hidden layers in the DL structure and the number of neurons in each hidden layer. If the selection is not reasonable, it not only increases the computational complexity of the method, but also reduces the diagnostic accuracy of the algorithm. Typically, the number of hidden layers is given according to different problems combined with human experience. Although it can achieve good results, it is not necessarily the best. Although, the number of neurons in each layer is currently selected by an optimization algorithm, but the optimization effect needs to be further improved.
(b) The training and testing sets used in the current methods are obtained from the same dataset, and the distribution of data structures in these two datasets is generally consistent. When a trained diagnosis model is used to diagnose the fault of a similar component with a different distribution of data, its diagnostic accuracy generally decreases.
(c) Most DL learning processes are essential black-box problems. It is impossible to clearly indicate the type of learning process inside. Only by clearly understanding the DL learning process can we improve some of the deficiencies in the learning process, so that the algorithm can improve the diagnostic accuracy by reducing the computational complexity. Although some visualization tools can show the learning effect of DL through each hidden layer, its effect still needs to be further improved.
(d) Although DL has high diagnostic accuracy in fault diagnosis, there is still a risk of error diagnosis. Consequently, multimodal information should be considered in data learning, which could further improve the accuracy of fault diagnosis.

## V. TRANSFER LEARNING (TL)

In previous methods, training sets and testing sets for model training and diagnosis are usually under the same working conditions. Moreover, they are subject to the same data distribution and have the same feature space, which is quite different from the actual situation. Because the data acquired in WT are usually nonlinear and non-stationary, the distribution of data for fault diagnosis and the distribution of training information for model training are frequently inconsistent, even though DL has achieved high diagnostic accuracy and is robust in troubleshooting. However, when the trained DL diagnostic network is used to diagnose a fault problem that has never been learned, the diagnostic accuracy is significantly reduced. To solve the above issues, researchers introduced TL into the fault diagnosis of WT. A basic fault classification schematic diagram of the TL is shown in Fig.7.

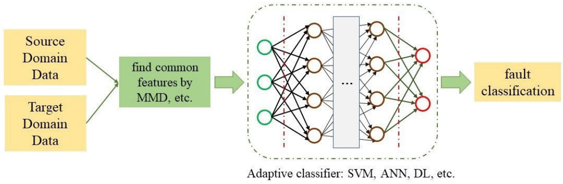

Source Domain Data
Target Domain Data
find common features by MMD, etc.
...
fault classification
Adaptive classifier: SVM, ANN, DL, etc.

**FIGURE 7. Typical feature-based TL applied for fault classification.**

### A. THEORETICAL BASIS

The current TL methods mainly can be divided into four types:

(a) Instance-based TL. Part of the source domain data can still be used in the target domain by overweighting and training;
(b) Feature-based TL. Fault features from the source domain are encoded and transferred to the target domain to improve the learning effect;
(c) Parameter-based TL. The target and source domains share the same model parameters or the same a priori distribution;
(d) Relevant knowledge-based TL. The source and target domains have the same or similar data distributions. The trained fault model can be used to maintain high diagnostic accuracy for similar problems under different diagnostic conditions. Moreover, TL can diagnose mechanical components based on known fault diagnosis models that do not have sufficient fault samples.

### B. APPLICATIONS

#### 1) FEATURE-BASED TL

Kandaswamy *et al.* [127] introduced an approach to improve TL accuracy by reusing SDAE. Data were learned through SDAE to complete the fault diagnosis for different data distributions and various tasks. The experiments demonstrated that the TL of high-dimensional features cannot only reduce the computational complexity of fault diagnosis, but also achieve higher diagnostic accuracy. However, the problem of negative transfer in the TL process needs to be solved. Shen *et al.* [128] presented a bearing fault diagnosis based on SVD feature extraction and TL classification. Using auxiliary data, the target data and additional data were given different weights by the TrAdaBoost algorithm, and then the TL algorithm was used for fault diagnosis. In this process, the negative transfer was avoided by similarity judgments, and the accuracy of the diagnosis was improved while reducing the computational complexity. Tong *et al.* [129] introduced a bearing fault diagnosis method under variable working conditions based on domain adaptation using feature-based TL. The dataset of normal and faulty bearings was obtained through the fast Fourier transformation of raw vibration signals under various conditions. Then, the marginal and conditional distributions were reduced simultaneously between the training data and testing data by refining pseudo test labels based on the maximum mean discrepancy and domain-invariant clustering in a common space. Finally, a transferable feature representation for the training data and testing data was obtained. Experimental results showed that this method is superior to other traditional methods.In Ref. [130] a fault diagnosis method for rolling bearings based on a sparse denoising autoencoder (SDAE) for deep feature extraction combining TL was proposed to improve the accuracy of bearing fault diagnosis. The bearing vibration signal in the time domain was transformed into the frequency domain, and the joint geometrical and statistical alignment was introduced to deal with the deep feature samples to reduce the domain discrepancy both statistically and geometrically. Additionally, the k-nearest neighbor classification algorithm was used to complete the fault diagnosis of rolling bearings under variable working conditions. To solve the problem of gear fault diagnosis, the fault features under known and unknown working conditions were selected by means of transferring feature analysis to minimize the difference between domains, which ensures that the data features remain unchanged while the data scale is reduced [131]. The selected fault features were used for fault diagnosis using the SVM. The effectiveness of this method was illustrated by comparison with PCA, kernel PCA (KPCA), locally linear embedding (LLE) and factor analysis (FA). Wang *et al.* [132] presented a heterogeneous TL method based on stack sparse autoencoders (SSAE) for fault diagnosis. The data in the source and target domains were represented by heterogeneous features of different dimensions, which were sent to the same feature space through different AE. The similarity degree of the two domains was judged according to the center distance between the domains, and then the difference between the data was reduced by using the maximum mean error. The final results were diagnosed by SVM and performed better than the TML approaches when there was little labeled data in the target domain. Ren *et al.* [133] proposed a fault diagnosis method based on VMD multiscale permutation entropy (MPE) and feature-based TL for WT. Aimed at the problem that the source and target domains data belong to different working conditions, the proposed method reduces the difference in data distribution between the source and target domains by minimizing the covariance between them through a linear transformation matrix. The experiment showed that the proposed covariance alignment (COVAL) of fault features has higher accuracy in rolling bearing multi-state classification under variable working conditions as compared with other methods. Qian *et al.* [134] constructed a novel deep transfer network (DTN) for rotating machine fault diagnosis with working condition variations, which combines auto-balanced high-order Kullback-Leibler (AHKL), smooth conditional distribution alignment (SCDA), and weighted joint distribution alignment (WJDA). Extensive experimental evaluations through 18 TL cases demonstrated its validity, and further comparisons with the state of the arts also validated its superiority. In order to overcome the weaknesses of the Gaussian kernel-induced maximum mean discrepancy (GK-MMD), Yang *et al.* [135] proposed a distance metric called polynomial kernel-induced MMD (PK-MMD). Combined with PK-MMD, a diagnosis model was constructed to reuse diagnosis knowledge from one machine to the another. The proposed methods were verified by two TL cases, and the results showed that the PK-MMD-based diagnosis model presented better transfer results than other methods.

### 2) PARAMETER-BASED TL

In Ref. [136], a network fault diagnosis model was established by sampling the data of the source domain, and the same operation was performed on the target domain data, which were then used to update the established network diagnostic model parameters, finally, and a small number of samples were obtained from the target domain tested. The results indicated that the method can reduce the network training time and improve the network diagnostic accuracy under the premise of a modest number of samples. In Ref. [5], [137], [138], the CNN was combined with TL to complete the fault diagnosis under different speeds and loads. First, the two-dimensional signal obtained by processing the source domain signal was sent to the CNN for network training, which contains a measure that does not change owing to the change in the working state. The CNN was updated with the signal of the target domain to adapt to the fault diagnosis under the new working state. The experiments demonstrated that the accuracy of the method mentioned in Ref. [137] reached 99.8%. In Ref. [13], the two-dimensional signals obtained after processing were subjected to diagnosis and testing of the network through AlexNet-based TL CNN, and the learning process was visualized using t-SNE during network training. The experimental results showed that the fault diagnostic accuracy of this method was 99.89%, which was higher fault diagnosis accuracy than that of CNN, DAE, DBN, and SVM. In Ref. [14], a fault diagnosis model using the SVM based on known samples was established, when there was a change in the diagnostic data, the fault features of the data before and after the change were placed in the same feature space. Moreover, a parameter to determine the degree of dependence on the existing model was set according to different parameter values to update the kernel function matrix in the SVM until the model could adapt to the new data distribution. Experiments conducted on toy data and real datasets showed that the proposed method could adapt to data change, and it provides a good transition from the old detection rule to the new one, which is obtained using the new data set only when the number of samples gathered from that new one is large enough. In addition, some TL methods that combine parameter-based TL and feature-based TL have also been applied. These methods usually first modify some existing deep network models, then add a domain adaptation layer to the network, and finally perform joint training. In Ref. [139], a novel domain adversarial transfer network (DATN) was proposed to handle large distribution discrepancies across domains. First, hierarchical representations were learned from the source and target domains with two asymmetric encoder networks. Then, the network weights learned in the source tasks are transferred to improve the training on the target tasks. Finally, the difference between the source and target distributions is minimized by domain adversarial training. The experimental results on two fault datasets demonstrated that the proposed method achieves excellent accuracy which outperforms other algorithms. In Ref. [140], an optimal ensemble deep transfer network (OEDTN) is proposed for rolling bearing fault diagnosis with unlabeled data, which takes advantage of parameter TL, domain adaptation and ensemble learning. Experiments on three bearing test rigs were carried out, and the results showed that the proposed method was more effective than the existing methods.

### 3) INSTANCE-BASED TL

Wen *et al.* [141] presented a new deep TL method based on the SAE for fault diagnosis. For the fault diagnosis of different bearing conditions, an SAE fault diagnosis network based on the source domain data was first established, and then the discriminate penalty of the training data and the testing data was minimized by the maximum mean discrepancy to update the fault diagnosis network with the target data. The method obtained more potential features by using unmarked third-party data, and the diagnostic accuracy of the operation was 99.82% through experiments under different operating conditions. Moreover, the final diagnostic accuracy was proportional to the standard deviation of the data. To address the problem of sufficient labeled samples in the laboratory equipment and a small number of unlabeled samples from actual machines, a TL method for intelligent fault diagnosis was presented [142]. First, the labeled samples were utilized to train the domain-shared CNN model, and the multi-kernel maximum mean discrepancy was then applied to minimize the error in the learning characteristics of the laboratory data and actual data. Subsequently, a pseudo label is generated for the actual unlabeled data to diagnose the fault through the domain shared classifier, and the validity of the method was proved experimentally. Xie *et al.* [143] introduced a TL strategy for rotating machinery fault diagnosis based on cycle-consistent generative adversarial networks (GAN). The characteristic of the network was that it could generate new samples similar to the original data through the training process. When the distributed adaptive signals were unified in all states, the samples generated under different working conditions could be approximated successfully. Based on the fault classifier established for learning the known samples, cycle-consistent GANs were used to generate different samples for different working states and improve the established classifiers to adapt to the fault diagnosis under different working conditions.

## C. SUMMARY OF TL

Because WT operate under variable conditions, the speed of the main shaft and the load of the blades are continually changing. In the past, the data of the training and testing sets in the fault diagnosis method obey the same distribution, but this does not meet the actual working conditions. Applying TL as a fault diagnosis method ensures that the designed method still maintains a high diagnostic accuracy for fault diagnosis under variable working conditions. A summary of the applications of TL for WT fault diagnosis is presented in Table 4.

**TABLE 4.** The summary of applications of TL to WT fault diagnosis.

| Methodologies | Monitoring Components | References |
|---|---|---|
| Feature-based | Bearings,Gearbox | Kandaswamy et al. [127], Fei et al. [128], Tong et al. [129], Dong et al. [130], Wang et al. [131], Wang et al. [132], Ren et al. [133], Qian et al. [134], Yang et al. [135] |
| Parameter-based | Bearings,Gearbox | Zhang et al. [136], Hasan et al. [137], Hasan et al. [138], Chen et al. [139], Li et al. [140] |
| Instance-based | Bearings,Gearbox | Wen et al. [141], Yang et al. [142], Xie et al. [143] |

Moreover, for some new fault problems, in the case of a small sample using the previous diagnostic knowledge, it is possible to quickly adapt to changes in the data in a short time, thereby enabling fault diagnosis. However, in the TL process, negative transfer learning usually occurs. The existing methods often use the maximum average difference to suppress negative transfer learning, and the establishment of a third-party data can also avoid negative transfer learning and improve TL effects simultaneously. Because the collected data are usually unmarked for the newly created fault problem, this is not very helpful for the domain sharing classifier. The current method adds a pseudo label to the data, but the diagnostic accuracy of the method needs to be improved. The current TL method can only diagnose faults under different working conditions for the same fault problem and cannot diagnose faults between different machines. The realization of the transfer of fault diagnosis knowledge between machines is a research direction, and how to use the existing fault samples for fault diagnosis is also a research field when a new fault problem occurs without any fault samples.

## VI. DISCUSSIONS

This paper summarizes the applications of ML in the fault diagnosis of WT and analyzes the advantages and disadvantages of various methods as shown in the Appendix.

For TML, SVM and its variants are the most widely used methods. Because the classic SVM is only suitable for dealing with the two classification problems, the real fault diagnosis is a multi-classification problem. Hence, it is necessary to make some improvements to the classic SVM to meet the requirements of practical problems. In Ref. [24], [32], the rough neighborhood set, manifold learning, and PCA algorithms were used to select the most relevant fault features, which can reduce the calculation time and improve the fault accuracy in the process of fault diagnosis. Different methods were applied to optimize the kernel parameters and penalty factors in the SVM to speed up the fault diagnosis. LSSVM and RVM, as variants of SVM, both improve the calculation efficiency of SVM [6], [20], [27], [33]. In Ref. [34], it fuses heterogeneous information to solve the misalignment problem in WT, and the t-SNE method was utilized for dimension reduction. In Ref. [29], the fault features of the collected data were used to form a two-dimensional vector for model training. By selecting the appropriate penalty factor and kernel function parameters to optimize the SVM, the diagnosis model can have a higher diagnostic accuracy.

In terms of DT, the C4.5 algorithm was used in Ref. [35], [36] and the J48 algorithm was used in Ref. [37], [38] to diagnose the faults of WT. The Bayesian method can realize fault diagnosis by minimizing the risk of the sample conditions. In Ref. [40], three Bayesian diagnosis models based on SCADA were established and compared with other methods to demonstrate the effectiveness of Bayesian diagnosis. The problem of signal transmission is solved and the data after transmission is reconstructed using sparse Bayes combined with other methods [41]. In Ref. [44], the state transition probability matrix was obtained using the Bayes method for fault diagnosis. The results showed that the method can predict faults 33 days in advance on average. In Ref. [42], [43] and [46], [47], different types of Bayesian classifiers were constructed according to a conventional Bayesian splitter in combination with specific problems. However, the diagnostic networks need to know the prior probability when they are applied. As for the HMM, the prediction of the next time state is completed only by the data of the current state. In Ref. [50], the semi-hidden Markov model was used to learn the fault features, and the posterior probability was tested to determine whether there was a fault in WT. In Ref. [52], the influence of the outliers in the data was reduced, making the HMM more sensitive to the fault characteristics. HMM usually uses the posterior probability to diagnose faults, but the posterior probability cannot be directly measured. Therefore, it is generally obtained indirectly through Bayes' theorem.

The RF is established on the basis of DT, and the diagnosis is completed by learning the same fault problem with multiple learners. In Ref. [55], the attributes were optimized by the GA in Ref. [58], [59], the fault features were screened using the correlation coefficient and Gini index respectively. In Ref. [63], the high-dimensional features of the DBM were fused by RF to diagnose the fault. The experiment showed that the diagnostic accuracy of this method was 97.68%. However, the selection and optimization of the learners are very important for the accuracy of the diagnosis in this method. The CA uses the membership degree between sample and cluster center to diagnose faults. The FCM algorithm was utilized in Ref. [60], and kernel FCM was used in [61] to improve the diagnosis accuracy by optimizing the cluster center. In Ref. [59], [62], [63], and K-means CA was applied. In Ref. [62], K-means CA was applied to deal with the outliers in the data. In Ref. [63], the historical data obtained were grouped by the K-means CA to establish the corresponding model for fault diagnosis, but this method needs to optimize the selection of the clustering center.

For ANN, fault samples are learned by connecting the weights of the neurons and adjusting the thresholds of the neurons. As the most widely applied ANN, BPNN completes the construction of a fault diagnosis network by forward propagation and backward fine-tuning. In Ref. [69], [71], various optimization algorithms, and the LM algorithm were utilized to optimize the weight and threshold value in BPNN to improve the convergence speed and diagnosis accuracy. The defect that it is easy to converge to the local optimal value is overcome as well. ELM is an improved algorithm for BPNN, and the connection weights between the input layer and hidden layer and the threshold of the hidden layer neurons were set randomly and not updated. In Ref. [77], a dual-ELM diagnosis model was proposed, with one of the two ELMs used for counting and the other was used to determine the possible fault type. In Ref. [74], a hierarchical ELM was established, and the diagnostic accuracy of this method was improved by 5%-10% as compared with other methods. In Ref. [78], an online sequential extreme learning machine (OS-ELM) method was proposed to realize online fault diagnosis. Moreover, this method can adapt to the new fault diagnosis by increasing the number of hidden layer neurons. Compared with BPNN, ELM has improved the diagnosis effects comprehensively.

The RBF network has a strong learning ability and can realize global optimization, but the method requires a high-quality fault samples. SOM is a competitive learning network, and the output results are displayed in a two-dimensional plane. Therefore, this method can be used as a visual method. However, its computational complexity is high, and when there are new faults, the entire network needs to be reconstructed. ART is also a type of competitive learning network. It can diagnose faults by similarity between neurons, and it is also an incremental learning method that can learn new faults by increasing the number of neurons.

Regarding the DNN, fault feature acquisition and fault diagnosis are integrated to eliminate the defects of human experience on the selection of fault features. The CNN reduces the complexity of the fault diagnosis by convolution and pooling. Combining CWT with CNN results in a fault diagnosis accuracy of 99.11% [102], and combining DWT with CNN promoted the fault diagnosis accuracy to a level of 99.3% [10]. Moreover, the LNP algorithm was used to visualize the CNN learning process [97]. Ref. [96] reported an algorithm that can reduce the training time by 80% without affecting the diagnosis accuracy. These methods extend the application scope of CNN.

The DBN is composed of a series of RBM, which realizes forward training through a greedy algorithm, and then performs reverse fine-tuning for the weights of DBN. In Ref. [111], PSO was used to optimize the number of hidden neurons and the learning rate in the DBN. Improved sigmoid units were applied to solve the problem in which the gradient of the DBN disappears during the reverse fine-tuning process [112].

Similarly, SAE is made up of several AE stacks, and its training process is similar to that of the DBN. In Ref. [116], an SAE was proposed to extract data containing multiple noise levels. In Ref. [118], SAE was used to learn bearing vibration signals, and the diagnosis accuracy reached 98%. In Ref. [119], the high-dimensional fault features were fused by locality preserving projection (LPP) to improve the learning ability. The RNN can maintain the state of the system at the last moment while receiving the external inputs. As the most widely used network in RNN, LSTM networks have been used to solve the problem of gradients disappearance in the reverse fine-tuning stage [122]. In Ref. [126], the high-dimensional features obtained by LSTM were sent to the RF for fault diagnosis. Although the application of the DNN algorithm greatly improves the accuracy of fault diagnosis, at present the learning process of each stage of the DL method cannot be understood, which is actually a "black box" problem.

TL solves the defect in which the diagnosis accuracy decreases when the existing diagnostic methods are applied to other similar problems. In feature-based TL, some auxiliary data were used in Ref. [128] to avoid negative transfer during the learning process. In Ref. [129]–[132], the fault features of the source and target domains were sent to the same feature space using different methods, and the discrepancy between the source domain and the target domain was reduced by selecting appropriate feature variables for fault diagnosis. In the parameter-based TL, SVM and SDAE were utilized to establish the TL model in Ref. [11], [127], but a negative transfer learning problem occurs in both methods. In Ref. [4], [13], [137], [138], the CNN was used to learn the data of the source domain, and then the trained model was modified by the sample of the target domain so that the fault diagnosis cloud be performed in the new environment. The final diagnostic accuracy for bearing fault diagnosis was 99.8% [137]. For the instance-based TL method, the laboratory fault samples were first learned to establish a diagnostic network, and then the TL was used to generate a pseudo label for the unmarked target domain data for fault diagnosis [142]. In Ref. [143], cycle-consistent GANs have been to create more target domain samples for fault diagnosis. In Ref. [141], third-party unlabeled data were used to obtain more potential features to further improve the diagnostic accuracy. The experimental results showed that the diagnostic accuracy of the method was 99.82%.

Although TL can expand the application scope of existing knowledge, there is occasionally a negative transfer phenomenon in the process of TL, and the application of the existing fault diagnosis model to other machinery fault diagnosis needs further research.

Normally, condition monitoring of WT can be considered as a pattern recognition problem that consists of three phases namely, feature extraction, feature selection, and feature classification [38]. In this review, different methods related to the phases of fault diagnosis are shown in Fig. 8. Most TML and ANN methods focus on the feature classification phases. They usually combine with traditional signal processing approaches, such as EMD and wavelet transform to extract features and then send them to TML or ANN for faults classification. DL and TL can include the phases from feature extraction to feature classification so that this type of methods can achieve fault diagnosis from end to end.

Following the development of TML, ANN, DL and TL, the intelligent fault diagnosis methods have released the human labor and demonstrated the potential of automatically recognizing WT fault patterns with great accuracy. From TML to TL, the performance of intelligent fault diagnosis methods has been improved. However, several challenges still exist in current studies.

(a) As is well known, the high diagnostic accuracy of the ML-based methods relies on the huge amount of data and the thousand times of training. The training process usually requires a long period of time to obtain satisfactory performance if it starts from zero. However, in real industrial scenarios, the diagnosis method is required to be active as soon as possible. Therefore, a method to reduce the calculation time of the training of the ML-based fault diagnosis models needs to be studied in the future. For this issue, two research interests are recommended for future research. 1) It is necessary to propose a method that can determine the trade-off between the complexity of the model, such as the number of hidden layers in the network and the number of neurons in each hidden layer, and the time consumption of the training. 2) The pre-trained model in the experimental environment could be leveraged to release the burden caused by training from zero.

(b) With the development of ML technologies, the DL-based fault diagnosis methods have become the mainstream of intelligent diagnostic methodologies. Although the unparalleled feature extraction ability of DL and its end-to-end diagnosis process have indeed facilitated intelligent algorithms for

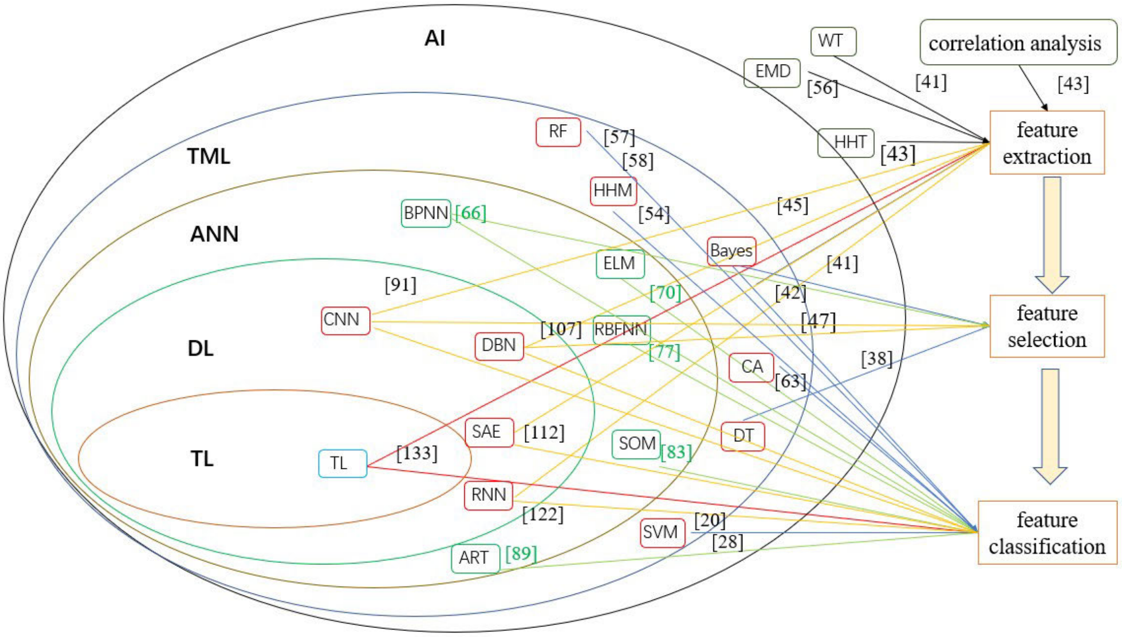


**FIGURE 8.** The distribution diagram of the ML methods applied for WT fault diagnosis.

WT fault diagnosis, the drawback of the black box for deep hierarchical networks still impedes the application of DL-based diagnosis methods in real-world scenarios. The diagnosis results of DL are mainly based on the estimation of sufficient data and thousands of experimental trials rather than a strictly theoretical background. Therefore, the diagnosis results of DL lack strict prior constraints and the physical meaning of the parameter and the extracted features of those models can hardly be explained. However, the diagnosis of WT is a serious process that requires diagnostic methods to be highly reliable. Improving the interpretability of DL-based methods is an important future research direction.

(c) For different types of WT, their internal structures are usually different, and they usually work under time-varying conditions. Therefore, the WT diagnosis models constructed by traditional ML, ANN and DL could hardly be well leveraged in all machines because of the different data distributions. To bridge this gap, TL provides intelligent fault diagnosis with a promising way. However, the inherent issue of negative transfer has rarely been discussed in the current work for WT fault diagnosis. Because of the high cost of data collection and the harsh operating environment, it is difficult to organize sufficient data in all domains. Therefore, the design of a specific TL-based paradigm for WT diagnosis that can reduce the negative transfer caused by limited data in different domains is worth investigating in future research.

(d) The working conditions of WT vary in many ways, so the degradation mode of the gearbox is not fixed. Studying and classifying the infinite working conditions as the limited degradation mode is a potential way to solve this type of problem. Wind speed, load, current and other information can reflect changes in working conditions. Identifying the degradation mode dynamically by multi-sensor self-mapping under changing working conditions still needs to be further investigated.

## VII. CONCLUSION

Fault diagnosis for WT is critical in terms of improving the economic profits of wind farms and maintaining safe operation. ML has shown a great potential for fault diagnosis of the key components of WT, such as gearbox, bearing, blade, etc. Moreover, the condition monitoring of WT has the characteristics of big data, and is thus more suitable for ML-based fault diagnosis.

In this paper, we review several typical ML methods for the fault diagnosis of WT from the perspectives of theoretical fundamentals and practical applications. By introducing TML, such as SVM, DT and HMM, the diagnostic approaches are able to automatically recognize the fault patterns of WT. However, these methods still rely on artificial feature extraction, which requires significant human labor and expert knowledge. To this end, the ANN is utilized to adaptively learn the health states of WT because of its multi-layer perceptron architecture, whereas the shallow network of ANN limits its performance. With the rapid development of ML, the DL-based fault diagnosis methods have become mainstream owing to their end-to-end diagnosis process and the performance with high accuracy. It should be noted that the successes of DL-based diagnosis models are subject to the situation of sufficient labeled samples which is probably impractical in real-world scenarios. To bridge

**TABLE 5.** Comparison of different ML methods for WT fault diagnosis.

| | Methods | Articles | Input data | Advantage | Disadvantage | Monitoring components |
|---|---|---|---|---|---|---|
| TML | SVM | [17]–[34] | Vibration signal, SCADA | Higher diagnostic accuracy with less and simple sampleS;<br>Good global optimization and generalization capabilities. | Rely on the selection of kernel function and penalty factor;<br>Converge to a local minimum easily. | Bearings, Gears, Blades |
| | DT | [35]–[39] | Vibration signal, generator current signal | Can deepen the understanding of faults;<br>Does not need to do much processing on the data;<br>It is not sensitive to information loss. | Converge to a local minimum easily;<br>Risk of overfitting;<br>Cannot establish an active online diagnostic network. | Gears, Generation system |
| | Bayes | [40]–[47] | Vibration signal, SCADA | Can make the fault diagnosis fast and straightforward;<br>The computational complexity does not increase significantly for multi-classification problems;<br>It is not sensitive to data loss. | It is sensitive to error category;<br>It is sensitive to data attributes;<br>Rely on hypothetical model. | Gearbox, Blades, Bearings |
| | HMM | [48]–[54] | Vibration signal | Can indicate the reasoning process. | The historical data cannot be fully utilized;<br>The topology in the diagnostic model is sometimes ambiguous. | Bearings |
| | RF | [55]–[59] | Vibration signal | Robust;<br>It is not sensitive to the outliers in the data. | Training process is complicated;<br>The calculation amount is large;<br>Time-consuming. | Bearings, Bearings |
| | CA | [60]–[64] | Vibration signal, SCADA | Can manage massive quantities of data with high efficiency. | The number of categories distinguished needs to be given in advance;<br>Depend on the selection of the initial cluster center. | Bearings, Gearbox |
| ANN | BPNN | [68]–[71] | Vibration signal | Self-learning and self-adjustment;<br>Fault tolerance ability. | Convergence speed is slow;<br>It is easy to converge to the local optimal value;<br>A certain risk of overfitting. | Gearbox, Transmission chain, generator fault |
| | ELM | [72]–[78] | Vibration signal, SCADA | Fast learning speed;<br>Can adapt to new situations;<br>Better generalization ability. | Only one layer of hidden neurons;<br>The learning ability is limited. | Gearbox, Transmission chain |
| | RBFNN | [79], [80] | Vibration signal | Strong nonlinear fitting ability;<br>Fast convergence speed;<br>Can achieve global optimization;<br>Better generalization ability. | Rely on the quality of the learning data and the selection of samples;<br>Information loss. | Blades, Actuators |
| | SOM | [81]–[87] | Vibration signal, SCADA | Visualization;<br>Implementation structure is relatively simple. | Calculating complexity is relatively high. | Bearings, Gearbox |
| | ART | [88]–[92] | Vibration signal | Can learn new problems while retaining previous learning experiences;<br>Do not rely on prior knowledge. | Information loss. | Bearings, Gearbox |
| DL | CNN | [94]–[105] | Vibration signal | Reduces the complexity of the network model;<br>Can avoid overfitting of data;<br>Better generalization ability;<br>Robust. | A large amount of data are required;<br>High computational complexity;<br>Can only process input data with a fixed length;<br>Information loss. | Bearings, Gearbox |
| | DBN | [106]–[112] | Vibration signal, SCADA | Does not depend on a large amount of signal processing technologies and experience;<br>Strong versatility and adaptability;<br>Can process high-dimensional, nonlinear data. | Only capable of processing one-dimensional data;<br>Large calculation time. | Gearbox |
| | SAE | [113]–[119] | Vibration signal | Do not need large amount of data;<br>Can overcome the gradient diffusion problem. | Training time is long;<br>Risk of overfitting;<br>No clear rules to measure the selected results. | Bearings, Gearbox |
| | RNN | [120]–[126] | Vibration signal | Possible to diagnose slow-developing faults;<br>Can solve the problem of gradient disappearance. | No clear rules on how to choose the optimal number of hidden neurons. | Bearings, Gearbox |
| TL | TL | [127]–[143] | Vibration signal, SCADA | Maintain high diagnostic accuracy for fault diagnosis under variable working conditions;<br>Adaptive for new fault problems. | Negative transfer learning may occur; | Bearings, Gearbox |

this gap, TL is introduced to WT fault diagnosis mission by transferring the diagnosis knowledge gained from one task to another. Finally, this paper discusses the challenges of the current WT fault diagnosis and provides several potential prospects for future research. We believe that this review will provide a comprehensive reference to related researchers.

## APPENDIX

See Table 5.

## REFERENCES

[1] *REmap: Roadmap for a Renewable Energy Future, 2016 Edition*, IRENA Agency, Abu Dhabi, UAE, 2016.

[2] A. S. Brouwer, M. van den Broek, Ö. Özdemir, P. Koutstaal, and A. Faaij, "Business case uncertainty of power plants in future energy systems with wind power," *Energy Policy*, vol. 89, pp. 237–256, Feb. 2016.

[3] N. Beganovic and D. Söffker, "Structural health management utilization for lifetime prognosis and advanced control strategy deployment of wind turbines: An overview and outlook concerning actual methods, tools, and obtained results," *Renew. Sustain. Energy Rev.*, vol. 64, pp. 68–83, Oct. 2016.

[4] P. Cao, S. Zhang, and J. Tang, "Preprocessing-free gear fault diagnosis using small datasets with deep convolutional neural network-based transfer learning," *IEEE Access*, vol. 6, pp. 26241–26253, 2017.

[5] P. Ding, H. Wang, W. Bao, and R. Hong, "HYGP-MSAM based model for slewing bearing residual useful life prediction," *Measurement*, vol. 141, pp. 162–175, Jul. 2019.

[6] M. Heidari, H. Homaei, H. Golestanian, and A. Heidari, "Fault diagnosis of gearboxes using wavelet support vector machine, least square support vector machine and wavelet packet transform," *J. Vibroeng.*, vol. 18, pp. 860–875, Mar. 2016.

[7] R. Liu, B. Yang, E. Zio, and X. Chen, "Artificial intelligence for fault diagnosis of rotating machinery: A review," *Mech. Syst. Signal Process.*, vol. 108, pp. 33–47, Aug. 2018.

[8] Z. Sun, H. Sun, and J. Zhang, "Multistep wind speed and wind power prediction based on a predictive deep belief network and an optimized random forest," *Math. Problems Eng.*, vol. 2018, pp. 1–15, Jul. 2018.

[9] A. Kusiak, Z. Zhang, and A. Verma, "Prediction, operations, and condition monitoring in wind energy," *Energy*, vol. 60, pp. 1–12, Oct. 2013.

[10] A. Stetco, F. Dinmohammadi, X. Zhao, V. Robu, D. Flynn, M. Barnes, J. Keane, and G. Nenadic, "Machine learning methods for wind turbine condition monitoring: A review," *Renew. Energy*, vol. 133, pp. 620–635, Apr. 2019.

[11] W. Y. Liu, "A review on wind turbine noise mechanism and de-noising techniques," *Renew. Energy*, vol. 108, pp. 311–320, Aug. 2017.

[12] W. Y. Liu, B. P. Tang, J. G. Han, X. N. Lu, N. N. Hu, and Z. Z. He, "The structure healthy condition monitoring and fault diagnosis methods in wind turbines: A review," *Renew. Sustain. Energy Rev.*, vol. 44, pp. 466–472, Apr. 2015.

[13] R. X. Chen, X. Huang, L. X. Yang, X. Y. Xu, X. Zhang, and Y. Zhang, "Intelligent fault diagnosis method of planetary gearboxes based on convolution neural network and discrete wavelet transform," *Comput. Ind.*, vol. 106, pp. 48–59, Apr. 2019.

[14] Y. Xue and P. Beauseroy, "Transfer learning for one class SVM adaptation to limited data distribution change," *Pattern Recognit. Lett.*, vol. 100, pp. 117–123, Dec. 2017.

[15] A. P. Marugán, F. P. G. Márquez, J. M. P. Perez, and D. Ruiz-Hernández, "A survey of artificial neural network in wind energy systems," *Appl. Energy*, vol. 228, pp. 1822–1836, Oct. 2018.

[16] P. Ma, H. Zhang, W. Fan, C. Wang, G. Wen, and X. Zhang, "A novel bearing fault diagnosis method based on 2D image representation and transfer learning-convolutional neural network," *Meas. Sci. Technol.*, vol. 30, no. 5, May 2019, Art. no. 055402.

[17] Y. Li, M. Xu, H. Zhao, and W. Huang, "Hierarchical fuzzy entropy and improved support vector machine based binary tree approach for rolling bearing fault diagnosis," *Mechanism Mach. Theory*, vol. 98, pp. 114–132, Apr. 2016.

[18] L. Wenyi, W. Zhenfeng, H. Jiguang, and W. Guangfeng, "Wind turbine fault diagnosis method based on diagonal spectrum and clustering binary tree SVM," *Renew. Energy*, vol. 50, pp. 1–6, Feb. 2013.

[19] N. Li, R. Zhou, Q. Hu, and X. Liu, "Mechanical fault diagnosis based on redundant second generation wavelet packet transform, neighborhood rough set and support vector machine," *Mech. Syst. Signal Process.*, vol. 28, pp. 608–621, Apr. 2012.

[20] A. Agasthian, R. Pamula, and L. A. Kumaraswamidhas, "Fault classification and detection in wind turbine using cuckoo-optimized support vector machine," *Neural Comput. Appl.*, vol. 31, no. 5, pp. 1503–1511, May 2019.

[21] C. Lihui, L. Yang, and Z. Donghua, "Fault diagnosis of the planetary gearbox based on ssDAG-SVM," *IFAC-PapersOnLine*, vol. 51, no. 24, pp. 263–267, 2018.

[22] F. Cheng, Y. Peng, L. Qu, and W. Qiao, "Current-based fault detection and identification for wind turbine drivetrain gearboxes," *IEEE Trans. Ind. Appl.*, vol. 53, no. 2, pp. 878–887, Mar. 2017.

[23] C. L. Liu and W. X. Qi, "Research on fault diagnosis method of wind turbine based on wavelet analysis and LS-SVM," *Adv. Mater. Res.*, vols. 724–725, pp. 593–597, Aug. 2013.

[24] B. Tang, T. Song, F. Li, and L. Deng, "Fault diagnosis for a wind turbine transmission system based on manifold learning and Shannon wavelet support vector machine," *Renew. Energy*, vol. 62, pp. 1–9, Feb. 2014.

[25] X. Li, W. Yao, X. Yang, and J. Wang, "Bearings fault diagnosis based on wavelet analysis and support vector machine," in *Proc. Int. Conf. Chem., Mater. Food Eng.*, 2015, pp. 896–899.

[26] Q. W. Gao, W. Y. Liu, B. P. Tang, and G. J. Li, "A novel wind turbine fault diagnosis method based on intergral extension load mean decomposition multiscale entropy and least squares support vector machine," *Renew. Energ.*, vol. 116, pp. 169–175, Feb. 2018.

[27] X. Zhang and J. Zhou, "Multi-fault diagnosis for rolling element bearings based on ensemble empirical mode decomposition and optimized support vector machines," *Mech. Syst. Signal Process.*, vol. 41, nos. 1–2, pp. 127–140, Dec. 2013.

[28] Q. Xiong, Y. Xu, Y. Peng, W. Zhang, Y. Li, and L. Tang, "Low-speed rolling bearing fault diagnosis based on EMD denoising and parameter estimate with alpha stable distribution," *J. Mech. Sci. Technol.*, vol. 31, no. 4, pp. 1587–1601, 2017.

[29] B. Xu, H. Li, F. Zhou, B. Yan, Y. Liu, and Y. Ma, "Fault diagnosis of variable load bearing based on quantum chaotic fruit fly VMD and variational RVM," *Shock Vib.*, vol. 2019, pp. 1–20, Jan. 2019.

[30] X. L. An, D. X. Jiang, S. H. Li, and J. Chen, "Fault diagnosis of direct-drive wind turbine based on support vector machine," *J. Phys., Conf. Ser.*, vol. 305, Jul. 2011, Art. no. 012030.

[31] J. Hang, J. Zhang, and M. Cheng, "Application of multi-class fuzzy support vector machine classifier for fault diagnosis of wind turbine," *Fuzzy Sets Syst.*, vol. 297, pp. 128–140, Aug. 2016.

[32] H. Malik and S. Mishra, "Proximal support vector machine (PSVM) based imbalance fault diagnosis of wind turbine using generator current signals," *Energy Proc.*, vol. 90, pp. 593–603, Dec. 2016.

[33] Y. Xiao, Y. Hong, X. Chen, and W. Chen, "The application of dual-tree complex wavelet transform (DTCWT) energy entropy in misalignment fault diagnosis of doubly-fed wind turbine (DFWT)," *Entropy*, vol. 19, no. 11, p. 587, 2017.

[34] Y. Xiao, Y. Wang, and Z. Ding, "The application of heterogeneous information fusion in misalignment fault diagnosis of wind turbines," *Energies*, vol. 11, no. 7, p. 1655, Jun. 2018.

[35] H. Wang, A. Peng, and X. Wang, "A fast fault diagnosis method for wind turbine generator system based on rough set-decision tree," in *Proc. 2nd Int. Conf. Artif. Intell., Manage. Sci. Electron. Commerce*, 2011, pp. 3630–3633.

[36] I. Vamsi, G. R. Sabareesh, and P. K. Penumakala, "Comparison of condition monitoring techniques in assessing fault severity for a wind turbine gearbox under non-stationary loading," *Mech. Syst. Signal Process.*, vol. 124, pp. 1–20, Jun. 2019.

[37] H. Malik and S. Mishra, "Application of fuzzy Q learning (FQL) technique to wind turbine imbalance fault identification using generator current signals," in *Proc. IEEE 7th Power India Int. Conf. (PIICON)*, Nov. 2016, pp. 1–6.

[38] A. Joshuva and V. Sugumaran, "A data driven approach for condition monitoring of wind turbine blade using vibration signals through best-first tree algorithm and functional trees algorithm: A comparative study," *ISA Trans.*, vol. 67, pp. 160–172, Mar. 2017.

[39] H. Liu, X.-H. Dong, Z.-L. Yang, and K. Zheng, "The application of intelligent fuzzy inference to the fault diagnosis in pitch-controlled system," *Energy Proc.*, vol. 16, pp. 1839–1844, Jan. 2012.

[40] Z. Song, Z. Zhang, Y. Jiang, and J. Zhu, "Wind turbine health state monitoring based on a Bayesian data-driven approach," *Renew. Energy*, vol. 125, pp. 172–181, Sep. 2018.

[41] Q. Li, W. Hu, E. Peng, and S. Liang, "Multichannel signals reconstruction based on tunable Q-factor wavelet transform-morphological component analysis and sparse Bayesian iteration for rotating machines," *Entropy*, vol. 20, no. 4, p. 263, Apr. 2018.

[42] J. Yu, M. Bai, G. Wang, and X. Shi, "Fault diagnosis of planetary gearbox with incomplete information using assignment reduction and flexible naive Bayesian classifier," *J. Mech. Sci. Technol.*, vol. 32, no. 1, pp. 37–47, Jan. 2018.

[43] J.-H. Zhong, J. Zhang, J. Liang, and H. Wang, "Multi-fault rapid diagnosis for wind turbine gearbox using sparse Bayesian extreme learning machine," *IEEE Access*, vol. 7, pp. 773–781, 2018.

[44] J. Herp, M. H. Ramezani, M. Bach-Andersen, N. L. Pedersen, and E. S. Nadimi, "Bayesian state prediction of wind turbine bearing failure," *Renew. Energy*, vol. 116, pp. 164–172, Feb. 2017.

[45] D. Wang, "An extension of the infograms to novel Bayesian inference for bearing fault feature identification," *Mech. Syst. Signal Process.*, vol. 80, pp. 19–30, Dec. 2016.

[46] K. Li, Q. Zhang, K. Wang, P. Chen, and H. Wang, "Intelligent condition diagnosis method based on adaptive statistic test filter and diagnostic Bayesian network," *Sensors*, vol. 16, no. 1, p. 76, Jan. 2016.

[47] J. Yu, B. Ding, and Y. He, "Rolling bearing fault diagnosis based on mean multigranulation decision-theoretic rough set and non-naive Bayesian classifier," *J. Mech. Sci. Technol.*, vol. 32, no. 11, pp. 5201–5211, Nov. 2018.

[48] F. P. G. Márquez, A. M. Tobias, J. M. P. Pérez, and M. Papaelias, "Condition monitoring of wind turbines: Techniques and methods," *Renew. Energy*, vol. 46, pp. 169–178, Oct. 2012.

[49] M. S. Kan, A. C. C. Tan, and J. Mathew, "A review on prognostic techniques for non-stationary and non-linear rotating systems," *Mech. Syst. Signal Process.*, vols. 62–63, pp. 1–20, Oct. 2015.

[50] X. Li, V. Makis, H. Zuo, and J. Cai, "Optimal Bayesian control policy for gear shaft fault detection using hidden semi-Markov model," *Comput. Ind. Eng.*, vol. 119, pp. 21–35, May 2018.

[51] S.-H. Shin, S. Kim, and Y.-H. Seo, "Development of a fault monitoring technique for wind turbines using a hidden Markov model," *Sensors*, vol. 18, no. 6, p. 1790, Jun. 2018.

[52] Y. Zhao, Y. Liu, and R. Wang, "Fuzzy scalar quantisation based on hidden Markov model and application in fault diagnosis of wind turbine," *J. Eng.*, vol. 2017, no. 14, pp. 2685–2689, Jan. 2017.

[53] Y. Gao, F. Villecco, M. Li, and W. Song, "Multi-scale permutation entropy based on improved LMD and HMM for rolling bearing diagnosis," *Entropy*, vol. 19, no. 4, p. 176, 2017.

[54] J. Liu, Y. Hu, B. Wu, Y. Wang, and F. Xie, "A hybrid generalized hidden Markov model-based condition monitoring approach for rolling bearings," *Sensors*, vol. 17, no. 5, p. 1143, May 2017.

[55] H. Gan and B. Jiao, "Fault diagnosis of wind Turbine's gearbox based on improved GA random forest classifier," *DEStech Trans. Eng. Technol. Res.*, pp. 206–210, Sep. 2018, doi: 10.12783/dtetr/amee2018/2532.

[56] C. Li, R.-V. Sanchez, G. Zurita, M. Cerrada, D. Cabrera, and R. E. Vásquez, "Gearbox fault diagnosis based on deep random forest fusion of acoustic and vibratory signals," *Mech. Syst. Signal Process.*, vols. 76–77, pp. 283–293, Aug. 2016.

[57] T. Han and D. Jiang, "Rolling bearing fault diagnostic method based on VMD-AR model and random forest classifier," *Shock Vib.*, vol. 2016, pp. 1–11, Jun. 2016.

[58] X. Qin, Q. Li, X. Dong, and S. Lv, "The fault diagnosis of rolling bearing based on ensemble empirical mode decomposition and random forest," *Shock Vib.*, vol. 2017, pp. 1–9, Aug. 2017.

[59] R. Jia, F. Ma, J. Dang, G. Liu, and H. Zhang, "Research on multidomain fault diagnosis of large wind turbines under complex environment," *Complexity*, vol. 2018, pp. 1–13, Jul. 2018.

[60] C. Li, M. Cerrada, D. Cabrera, R. V. Sanchez, F. Pacheco, G. Ulutagay, and J. V. De Oliveira, "Some preliminary results on the comparison of FCM, GK, FCMFP and FN-DBSCAN for bearing fault diagnosis," in *Proc. Int. Conf. Sens., Diag., Prognostics, Control (SDPC)*, Aug. 2017, pp. 41–46.

[61] B. Wu and X. Li, "Fault diagnosis method based on kernel fuzzy C-means clustering with gravitational search algorithm," in *Proc. IEEE 7th Data Driven Control Learn. Syst. Conf. (DDCLS)*, May 2018, pp. 235–239.

[62] H.-H. Yang, M.-L. Huang, and S.-W. Yang, "Integrating auto-associative neural networks with hotelling T2 control charts for wind turbine fault detection," *Energies*, vol. 8, no. 10, pp. 12100–12115, 2015.

[63] J. Zhang, N. Jiang, H. Li, and N. Li, "Online health assessment of wind turbine based on operational condition recognition," *Trans. Inst. Meas. Control*, vol. 41, no. 10, pp. 2970–2981, Jun. 2019.

[64] L. Song and R. Yan, "Bearing fault diagnosis based on cluster-contraction stage-wise orthogonal-matching-pursuit," *Measurement*, vol. 140, pp. 240–253, Jul. 2019.

[65] C. Lu, C.-W. Fei, Y.-W. Feng, Y.-J. Zhao, X.-W. Dong, and Y.-S. Choy, "Probabilistic analyses of structural dynamic response with modified kriging-based moving extremum framework," *Eng. Failure Anal.*, vol. 125, Jul. 2021, Art. no. 105398.

[66] C. Lu, C.-W. Fei, H.-T. Liu, H. Li, and L.-Q. An, "Moving extremum surrogate modeling strategy for dynamic reliability estimation of turbine blisk with multi-physics fields," *Aerosp. Sci. Technol.*, vol. 106, Nov. 2020, Art. no. 106112.

[67] B. Keshtegar, M. Bagheri, C.-W. Fei, C. Lu, O. Taylan, and D.-K. Thai, "Multi-extremum-modified response basis model for nonlinear response prediction of dynamic turbine blisk," *Eng. Comput.*, no. 6, pp. 1–12, Jan. 2021, doi: 10.1007/s00366-020-01273-8.

[68] X. An, D. Jiang, and S. Li, "Application of back propagation neural network to fault diagnosis of direct-drive wind turbine," in *Proc. World Non-Grid-Connected Wind Power Energy Conf.*, Nov. 2010, pp. 1–5.

[69] L. Ju, D. Song, B. Shi, and Q. Zhao, "Fault predictive diagnosis of wind turbine based on LM arithmetic of artificial neural network theory," in *Proc. 7th Int. Conf. Natural Comput.*, Jul. 2011, pp. 575–579.

[70] Z. Wang and Q. Guo, "The diagnosis method for converter fault of the variable speed wind turbine based on the neural networks," in *Proc. 2nd Int. Conf. Innov. Comput., Inf. Control (ICICIC)*, Sep. 2007, p. 615.

[71] S. Han, J. Li, and Y. Liu, "Tabu search algorithm optimized ANN model for wind power prediction with NWP," *Energy Proc.*, vol. 12, no. 39, pp. 733–740, 2011.

[72] K. Li, L. Su, J. J. Wu, H. Q. Wang, and P. Chen, "A rolling bearing fault diagnosis method based on variational mode decomposition and an improved kernel extreme learning machine," *Appl. Sci.*, vol. 7, no. 10, p. 1004, Sep. 2017.

[73] M. F. Isham, M. S. Leong, M. H. Lim, and Z. A. Bin Ahmad, "Intelligent wind turbine gearbox diagnosis using VMDEA and ELM," *Wind Energy*, vol. 22, no. 6, pp. 813–833, Jun. 2019.

[74] Z.-X. Yang, X.-B. Wang, and J.-H. Zhong, "Representational learning for fault diagnosis of wind turbine equipment: A multi-layered extreme learning machines approach," *Energies*, vol. 9, no. 6, p. 379, 2016.

[75] P. Qian, D. Zhang, X. Tian, Y. Si, and L. Li, "A novel wind turbine condition monitoring method based on cloud computing," *Renew. Energy*, vol. 135, pp. 390–398, May 2019.

[76] B. Wu, L. Xi, S. Fan, and J. Zhan, "Fault diagnosis for wind turbine based on improved extreme learning machine," *J. Shanghai Jiaotong Univ., Science*, vol. 22, no. 4, pp. 466–473, Aug. 2017.

[77] X.-B. Wang, Z.-X. Yang, P. K. Wong, and C. Deng, "Novel paralleled extreme learning machine networks for fault diagnosis of wind turbine drivetrain," *Memetic Comput.*, vol. 11, no. 2, pp. 127–142, Jun. 2019.

[78] P. Qian, X. Ma, and D. Zhang, "Estimating health condition of the wind turbine drivetrain system," *Energies*, vol. 10, no. 10, p. 1583, Oct. 2017.

[79] A. Joshuva and V. Sugumaran, "A study of various blade fault conditions on a wind turbine using vibration signals through histogram features," *J. Eng. Sci. Technol.*, vol. 13, no. 1, pp. 102–121, 2018.

[80] V. Pashazadeh, F. R. Salmasi, and B. N. Araabi, "Data driven sensor and actuator fault detection and isolation in wind turbine using classifier fusion," *Renew. Energy*, vol. 116, pp. 99–106, Feb. 2017.

[81] Z. Zhemin, W. Tian, and L. Fenlan, "Fault diagnosis based on wavelet neural network," in *Proc. 5th Int. Conf. Intell. Comput. Technol. Autom.*, Jan. 2012, pp. 802–804.

[82] Y. Cunxiang and B. Hao, "The fault diagnosis of transformer based on the SOM neural network current," in *Proc. 5th Int. Conf. Measuring Technol. Mechatronics Autom.*, Jan. 2013, pp. 1178–1180.

[83] M. L. Fadda and A. Moussaoui, "Hybrid SOM–PCA method for modeling bearing faults detection and diagnosis," *J. Brazilian Soc. Mech. Sci. Eng.*, vol. 40, no. 5, p. 268, May 2018.

[84] O. Kramer, F. Gieseke, and B. Satzger, "Wind energy prediction and monitoring with neural computation," *Neurocomputing*, vol. 109, pp. 84–93, Jun. 2013.

[85] S. Wang, Y. Huang, L. Li, and C. Liu, "Wind turbines abnormality detection through analysis of wind farm power curves," *Measurement*, vol. 93, pp. 178–188, Nov. 2016.

[86] A. Gil, M. Sanz-Bobi, and M. Rodríguez-López, "Behavior anomaly indicators based on reference patterns—Application to the gearbox and electrical generator of a wind turbine," *Energies*, vol. 11, no. 1, p. 87, Jan. 2018.

[87] K. Gibert, P. Marti-Puig, J. Cusidó, and J. Solé-Casals, "Identifying health status of wind turbines by using self organizing maps and interpretation-oriented post-processing tools," *Energies*, vol. 11, no. 4, p. 723, Mar. 2018.

[88] D. Yang, H. Li, Y. Hu, J. Zhao, H. Xiao, and Y. Lan, "Vibration condition monitoring system for wind turbine bearings based on noise suppression with multi-point data fusion," *Renew. Energy*, vol. 92, pp. 104–116, Jul. 2016.

[89] Z. Li, Z. Ma, Y. Liu, W. Teng, and R. Jiang, "Crack fault detection for a gearbox using discrete wavelet transform and an adaptive resonance theory neural network," *J. Mech. Eng.*, vol. 61, no. 1, pp. 63–73, Jan. 2015.

[90] J. B. Ali, L. Saidi, S. Harrath, E. Bechhoefer, and M. Benbouzid, "Online automatic diagnosis of wind turbine bearings progressive degradations under real experimental conditions based on unsupervised machine learning," *Appl. Acoust.*, vol. 132, pp. 167–181, Mar. 2018.

[91] I. S. Lee, J. T. Kim, J. W. Lee, D. Y. Lee, and K. Y. Kim, "Model-based fault detection and isolation method using ART2 neural network," *Int. J. Intell. Syst.*, vol. 18, no. 10, pp. 1087–1100, Oct. 2003.

[92] X.-J. Wan, L. Liu, Z. Xu, and Z. Xu, "Gearbox fault diagnosis based on selective integrated soft competitive learning fuzzy adaptive resonance theory," *J. Comput. Inf. Sci. Eng.*, vol. 19, no. 1, Mar. 2019, Art. no. 011008.

[93] C.-W. Fei, H. Li, H.-T. Liu, C. Lu, L.-Q. An, L. Han, and Y.-J. Zhao, "Enhanced network learning model with intelligent operator for the motion reliability evaluation of flexible mechanism," *Aerosp. Sci. Technol.*, vol. 107, Dec. 2020, Art. no. 106342.

[94] F. Y. Zhou, L. P. Jin, and J. Dong, "Review of convolutional neural network," *Chin. J. Comput.*, vol. 40, no. 6, pp. 1229–1251, 2017, doi: 10.11897/SP.J.1016.2017.01229.

[95] Z. Chen, C. Li, and R.-V. Sanchez, "Gearbox fault identification and classification with convolutional neural networks," *Shock Vib.*, vol. 2015, pp. 1–10, Apr. 2015.

[96] R. P. Monteiro, M. Cerrada, D. R. Cabrera, R. V. Sánchez, and C. J. A. Bastos-Filho, "Using a support vector machine based decision stage to improve the fault diagnosis on gearboxes," *Comput. Intell. Neurosci.*, vol. 2019, pp. 1–13, Feb. 2019.

[97] J. Grezmak, P. Wang, C. Sun, and R. X. Gao, "Explainable convolutional neural network for gearbox fault diagnosis," *Proc. CIRP*, vol. 80, pp. 476–481, Jan. 2019.

[98] M. Xia, T. Li, L. Xu, L. Liu, and C. W. de Silva, "Fault diagnosis for rotating machinery using multiple sensors and convolutional neural networks," *IEEE/ASME Trans. Mechatronics*, vol. 23, no. 1, pp. 101–110, Feb. 2017.

[99] C. Sobie, C. Freitas, and M. Nicolai, "Simulation-driven machine learning: Bearing fault classification," *Mech. Syst. Signal Process.*, vol. 99, pp. 403–419, Jan. 2018.

[100] O. Janssens, V. Slavkovikj, B. Vervisch, K. Stockman, M. Loccufier, S. Verstockt, R. Van de Walle, and S. Van Hoecke, "Convolutional neural network based fault detection for rotating machinery," *J. Sound Vib.*, vol. 377, pp. 331–345, Sep. 2016.

[101] W. Fuan, J. Hongkai, S. Haidong, D. Wenjing, and W. Shuaipeng, "An adaptive deep convolutional neural network for rolling bearing fault diagnosis," *Meas. Sci. Technol.*, vol. 28, no. 9, Sep. 2017, Art. no. 095005.

[102] S. Guo, T. Yang, W. Gao, C. Zhang, and Y. Zhang, "An intelligent fault diagnosis method for bearings with variable rotating speed based on Pythagorean spatial pyramid pooling CNN," *Sensors*, vol. 18, no. 11, p. 3857, 2018.

[103] F. Chen, Z. Fu, and Z. Yang, "Wind power generation fault diagnosis based on deep learning model in Internet of Things (IoT) with clusters," *Cluster Comput.*, vol. 22, no. S6, pp. 14013–14025, Nov. 2019.

[104] Z. Xiaoxun, Z. Jianhong, H. Dongnan, and H. Zhonghe, "Research on mechanical rotor condition monitoring based on VCNN," *Energy Proc.*, vol. 158, pp. 6393–6398, Feb. 2019.

[105] Z. Yuan, L. Zhang, and L. Duan, "A novel fusion diagnosis method for rotor system fault based on deep learning and multi-sourced heterogeneous monitoring data," *Meas. Sci. Technol.*, vol. 29, no. 11, Nov. 2018, Art. no. 115005.

[106] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, "Deep learning and its applications to machine health monitoring," *Mech. Syst. Signal Process.*, vol. 115, pp. 213–237, Jan. 2019.

[107] G. Helbing and M. Ritter, "Deep Learning for fault detection in wind turbines," *Renew. Sustain. Energy Rev.*, vol. 98, pp. 189–198, Dec. 2018.

[108] D. Yu, Z. M. Chen, K. S. Xiahou, M. S. Li, T. Y. Ji, and Q. H. Wu, "A radically data-driven method for fault detection and diagnosis in wind turbines," *Int. J. Electr. Power Energy Syst.*, vol. 99, pp. 577–584, Jul. 2018.

[109] P. Tamilselvan and P. F. Wang, "Failure diagnosis using deep belief learning based health state classification," *Rel. Eng., Syst. Safety*, vol. 115, pp. 124–135, Jul. 2013.

[110] L. Xiuli, Z. Xueying, and W. Liyong, "Fault diagnosis method of wind turbine gearbox based on deep belief network and vibration signal," in *Proc. 57th Annu. Conf. Soc. Instrum. Control Eng. Jpn. (SICE)*, Sep. 2018, pp. 1699–1704.

[111] H. D. Shao, H. K. Jiang, X. Zhang, and M. G. Niu, "Rolling bearing fault diagnosis using an optimization deep belief network," *Meas. Sci. Technol.*, vol. 26, no. 11, Nov. 2015, Art. no. 115002.

[112] Y. Qin, X. Wang, and J. Zou, "The optimized deep belief networks with improved logistic sigmoid units and their application in fault diagnosis for planetary gearboxes of wind turbines," *IEEE Trans. Ind. Electron.*, vol. 66, no. 5, pp. 3814–3824, May 2019.

[113] F. Cheng, J. Wang, L. Qu, and Q. Wei, "Rotor-current-based fault diagnosis for DFIG wind turbine drivetrain gearboxes using frequency analysis and a deep classifier," *Ind. Appl. Soc. Meeting*, vol. 54, no. 2, pp. 1062–1071, 2017.

[114] Z. Wang, J. Wang, and Y. Wang, "An intelligent diagnosis scheme based on generative adversarial learning deep neural networks and its application to planetary gearbox fault pattern recognition," *Neurocomputing*, vol. 310, pp. 213–222, Oct. 2018.

[115] J. Yu, "A selective deep stacked denoising autoencoders ensemble with negative correlation learning for gearbox fault diagnosis," *Comput. Ind.*, vol. 108, pp. 62–72, Jun. 2019.

[116] G. Jiang, H. He, P. Xie, and Y. Tang, "Stacked multilevel-denoising autoencoders: A new representation learning approach for wind turbine gearbox fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 9, pp. 2391–2402, Sep. 2017.

[117] C. Lu, Z.-Y. Wang, W.-L. Qin, and J. Ma, "Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification," *Signal Process.*, vol. 130, pp. 377–388, Jan. 2017.

[118] F.-W. Qin, J. Bai, and W.-Q. Yuan, "Research on intelligent fault diagnosis of mechanical equipment based on sparse deep neural networks," *J. Vibroeng.*, vol. 19, no. 4, pp. 2439–2455, Jun. 2017.

[119] H. Shao, H. Jiang, F. Wang, and H. Zhao, "An enhancement deep feature fusion method for rotating machinery fault diagnosis," *Knowl.-Based Syst.*, vol. 119, pp. 200–220, Mar. 2016.

[120] S. Tang, S. Yuan, and Y. Zhu, "Deep learning-based intelligent fault diagnosis methods toward rotating machinery," *IEEE Access*, vol. 8, pp. 9335–9346, 2020.

[121] N. Talebi, M. A. Sadrnia, and A. Darabi, "Robust fault detection of wind energy conversion systems based on dynamic neural networks," *Comput. Intell. Neurosci.*, vol. 2014, no. 7, 2014, Art. no. 580972.

[122] P. Qian, X. Tian, J. Kanfoud, J. Lee, and T.-H. Gan, "A novel condition monitoring method of wind turbines based on long short-term memory neural network," *Energies*, vol. 12, no. 18, p. 3411, Sep. 2019.

[123] R. Yang, M. Huang, Q. Lu, and M. Zhong, "Rotating machinery fault diagnosis using long-short-term memory recurrent neural network," *IFAC-PapersOnLine*, vol. 51, no. 24, pp. 228–232, 2018.

[124] J. Lei, C. Liu, and D. Jiang, "Fault diagnosis of wind turbine based on long short-term memory networks," *Renew. Energy*, vol. 133, pp. 422–432, Apr. 2018.

[125] Z. Sun and H. Sun, "Health status assessment for wind turbine with recurrent neural networks," *Math. Problems Eng.*, vol. 2018, pp. 1–16, Dec. 2018.

[126] M. Li, D. Yu, Z. Chen, K. Xiahou, T. Ji, and Q. H. Wu, "A data-driven residual-based method for fault diagnosis and isolation in wind turbines," *IEEE Trans. Sustain. Energy*, vol. 10, no. 2, pp. 895–904, Apr. 2018.

[127] C. Kandaswamy, L. M. Silva, L. A. Alexandre, R. Sousa, J. M. Santos, and J. M. de Sá, "Improving transfer learning accuracy by reusing stacked denoising autoencoders," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2014, pp. 1380–1387.

[128] F. Shen, C. Chen, R. Yan, and R. X. Gao, "Bearing fault diagnosis based on SVD feature extraction and transfer learning classification," in *Proc. Prognostics Syst. Health Manage. Conf. (PHM)*, Oct. 2015, pp. 1–6.

[129] Z. Tong, W. Li, B. Zhang, F. Jiang, and G. Zhou, "Bearing fault diagnosis under variable working conditions based on domain adaptation using feature transfer learning," *IEEE Access*, vol. 6, pp. 76187–76197, 2018.

[130] S. Dong, K. He, and B. Tang, "The fault diagnosis method of rolling bearing under variable working conditions based on deep transfer learning," *J. Brazilian Soc. Mech. Sci. Eng.*, vol. 42, no. 11, pp. 1–13, Nov. 2020.

[131] J. Wang, J. Xie, L. Zhang, and L. Duan, "A factor analysis based transfer learning method for gearbox diagnosis under various operating conditions," in *Proc. Int. Symp. Flexible Autom. (ISFA)*, Aug. 2016, pp. 81–86.

[132] W. Chunfeng, L. Zheng, Z. Jun, and W. Wei, "Heterogeneous transfer learning based on stack sparse auto-encoders for fault diagnosis," in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2018, pp. 4277–4281.

[133] H. Ren, W. Liu, M. Shan, and X. Wang, "A new wind turbine health condition monitoring method based on VMD-MPE and feature-based transfer learning," *Measurement*, vol. 148, Dec. 2019, Art. no. 106906.

[134] W. Qian, S. Li, and X. Jiang, "Deep transfer network for rotating machine fault analysis," *Pattern Recognit.*, vol. 96, Dec. 2019, Art. no. 106993.

[135] B. Yang, Y. Lei, F. Jia, N. Li, and Z. Du, "A polynomial kernel induced distance metric to improve deep transfer learning for fault diagnosis of machines," *IEEE Trans. Ind. Electron.*, vol. 67, no. 11, pp. 9747–9757, Nov. 2020.

[136] R. Zhang, H. Tao, L. Wu, and Y. Guan, "Transfer learning with neural networks for bearing fault diagnosis in changing working conditions," *IEEE Access*, vol. 5, pp. 14347–14357, 2017.

[137] M. J. Hasan and J.-M. Kim, "Bearing fault diagnosis under variable rotational speeds using stockwell transform-based vibration imaging and transfer learning," *Appl. Sci.*, vol. 8, no. 12, p. 2357, Nov. 2018.

[138] J. J. Hasan, M. M. M. Islam, and J.-M. Kim, "Acoustic spectral imaging and transfer learning for reliable bearing fault diagnosis under variable speed conditions," *Measurement*, vol. 138, pp. 620–631, May 2019.

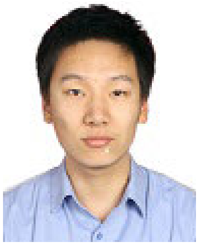
[139] Z. Chen, G. He, J. Li, Y. Liao, K. Gryllias, and W. Li, "Domain adversarial transfer network for cross-domain fault diagnosis of rotary machinery," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 11, pp. 8702–8712, Nov. 2020.

[140] X. Li, H. Jiang, R. Wang, and M. Niu, "Rolling bearing fault diagnosis using optimal ensemble deep transfer network," *Knowl.-Based Syst.*, vol. 213, Feb. 2021, Art. no. 106695.

[141] L. Wen, L. Gao, and X. Li, "A new deep transfer learning based on sparse auto-encoder for fault diagnosis," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 1, pp. 136–144, Jan. 2019.

[142] B. Yang, Y. Lei, F. Jia, and S. Xing, "A transfer learning method for intelligent fault diagnosis from laboratory machines to real-case machines," in *Proc. Int. Conf. Sensing, Diagnostics, Prognostics, Control (SDPC)*, Aug. 2018, pp. 35–40.

[143] Y. Xie and T. Zhang, "A transfer learning strategy for rotation machinery fault diagnosis based on cycle-consistent generative adversarial networks," in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2018, pp. 1309–1313.



**TONGDA SUN** was born in 1996. He received the B.S. degree in mechanical engineering from Chongqing University, China, in 2019. He is currently pursuing the master's degree with the Harbin Institute of Technology at Shenzhen, Shenzhen, China.

His research interests include the fault diagnosis and prognosis, condition monitoring of wind turbines, machine learning, and deep learning-based maintenance technologies for complex machinery systems.

**GANG YU** was born in 1969. He received the B.S. and master's degrees in mechanical engineering from the Dalian University of Technology, Dalian, China, in 1992 and 1995, respectively, and the Ph.D. degree from the University of Wisconsin–Milwaukee, USA.

Since 2005, he has been working with the Harbin Institute of Technology at Shenzhen, where he is currently an Associate Professor with the School of Mechanical Engineering and Automation. His current research interests include signal processing, fault prognosis and diagnosis, condition monitoring, intelligent maintenance, and service robots.

**MANG GAO** was born in 1990. He received the B.S. degree in petroleum engineering from Yangtze University (Jinzhou), China, in 2012. He is currently pursuing the master's degree with the Harbin Institute of Technology at Shenzhen, Shenzhen, China. His research interests include the fault diagnosis and prognosis, condition monitoring of wind turbines, and machine learning applied in complex machinery systems.

**LULU ZHAO** was born in 1996. He received the B.S. degree in vehicle engineering from the Harbin Institute of Technology (Weihai), China, in 2018. He is currently pursuing the master's degree with the Harbin Institute of Technology at Shenzhen, Shenzhen, China. His research interests include the fault diagnosis and prognosis, and transfer learning applied in complex machinery systems.

**CHEN BAI** was born in 1997. He received the B.S. degree from the Department of Mechanical Engineering, Dalian University of Technology, China, in 2019. He is currently pursuing the master's degree with the Harbin Institute of Technology, Shenzhen, China. His research interests include deep learning, transfer learning, and their application in fault diagnosis.

**WANQIAN YANG** was born in 1996. He received the B.S. degree from the Harbin Institute of Technology at Weihai, Weihai, China, in 2012. He is currently pursuing the master's degree with the Harbin Institute of Technology at Shenzhen, Shenzhen, China. His research interests include the fault diagnosis, predictive maintenance system for wind turbines, and the federated learning.