A review of recent advances in fault diagnosis based on deep neural networks

Rongyu Li
1576837338@qq.com
Xiamen University, Xiamen, Fujian, 361000,China

Abstract. Bearings are essential components in mechanical systems, supporting the rotation of various machine parts in motors, wind turbines, vehicles, and industrial robots, making their health critical for system performance and reliability. Traditional diagnosis methods, such as vibration and acoustic analysis, along with temperature monitoring, often demand expertise and may struggle to detect early faults. However, the introduction of deep learning technology has created new opportunities for more effective bearing fault diagnosis. The application of deep learning-based bearing fault diagnosis in the industrial sector has gained significant attention and multiple types of deep learning networks have already been successfully implemented. This paper aims to provide a clear review of bearing fault diagnosis based on deep learning algorithms. This essay focuses on two of the most popular deep learning networks, Autoencoder and Convolutional Neural Networks. Their mechanism and applications are analyzed based on essays and research paper related to the field of bearing fault diagnosis. Finally, conclusions are presented to summarize the current development and point out faced challenges and future trends of these deep learning networks. It is also expected that this narrative not only serves as a cogent overview of the contemporary fault diagnosis technologies but also provides convenience and inspiration for further study in this field.

Keywords: Deep Neural Networks, Fault Diagnosis, Autoencoder.

1. Introduction

Modern mechanical rotary machines are systems with great complexity which demands a large number of gears and bearings working together for target tasks [1]. Nevertheless, damage and failure of these machines are prone to happen due to the rough working environment, false operation, and long-term serves [2,3]. This is especially true for components called Rolling Element Bearings (REBs) whose broken may lead to nearly 50% of equipment failure cases and cause huge financial and time cost [4,5]. Thus, fault diagnosis technologies have to be applied in the industry to handle such problems. One of the significant focuses among them is the fault diagnosis of REBs, which has gained considerable attention from researchers. [6,7,8].

With the rapid development of computer technology, implementing computing technologies such as the compound fault diagnosis (CFD)[1] and the intelligent fault diagnosis (IFD) [9] have become one of the most popular methods of fault diagnosis of rotating machines. Compound fault diagnosis relies on predetermined rules and heuristics to identify single faults, often lacking the ability to account for complex interactions between faults [48]. In contrast, intelligent fault diagnosis employs advanced technologies like AI and machine learning to detect complex fault patterns and relationships within data. Intelligent fault diagnosis adapts and learns from new data, providing more accurate and comprehensive fault identification. It has gained more attention due to its ability to handle intricate system interactions, adapt to dynamic systems, offer real-time monitoring, and reduce downtime [49]. Advances in AI and ML have made IFD more accessible and practical, making it a desirable solution for improving operational efficiency and reducing maintenance costs in various industries [10]. When applying an IFD based on Machine Learning (ML), the examining of a REB's condition can be regarded as a pattern classification task. Basically, traditional ML method includes four steps.

First, the data collecting step acquires signals from sensor systems which reflect the health status of bearings in forms such as vibration signals [11] and electrical motor currents [13]. Then, with the help of tools such as Fourier Transform [14] and Wavelet Packet Transform [15], the process of feature extraction abstracts and identifies distinct signal that express information about whether the machine is working properly. The extracted feature set can then be compressed by using approaches like Principal Component Analysis [17] and Linear Discriminant analysis [19] which leaves only the most discriminant features to make the diagnosis more efficient. After the selected features are fed into classifiers such as k-Nearest Neighbor [20] and Artificial Neural Network [21,22], the type of bearing fault can be recognized.

Although bearing fault diagnosis based on machine learning, as mentioned earlier, has been widely applied for decades, this method does come with certain limitations. For one thing, Human expertise and signal processing techniques are required to design a feature extractor, but there's no common feature extraction process that can cater the need of every task, which means it is likely that a new feature extractor needs to be designed every time a new task appears [24]. For another thing, most machine learning algorithms only exploit shallow-structured architectures with a single layer of non-linear feature transformations, which makes them lack abilities of adapting to such features. Thus, a new and more effective method need to be found [25]. Another issue of the traditional approach based on ML is that it includes two stages. The first stage involves feature extraction, followed by input into the classifier. However, the features may not necessarily be suitable for the classifier, leading t5 o a decrease in performance [26]. The optimal approach is to consider both stages collaboratively, achieving end-to-end modeling.

In recent years, implementing Deep Learning (DL) algorithms has appeared to be one of the most desirable solutions for the problems mentioned in the last paragraph. DL is one of the most popular branches of ML whose concept first appeared in the 1980s. It is composed of layers of interconnected nodes (neurons). Each node processes information and passes it to the next layer. Deep networks have multiple hidden layers, allowing them to learn intricate representations. This is reason why Deep learning models are characterized by their depth. This depth enables them to capture hierarchical features and relationships in data. During the past few years, the techniques developed from deep learning research have already made lots of positive impact on a wide range of signal processing work [27]. It is currently widely used in fields such as neural network, graphical modeling [28], signal processing, and pattern recognition [30]. Deep learning networks such as Convolutional Neural Network (CNN), Stacked Autoencoder and Recurrent Neural Network have been applied successfully in areas including computer vision [31,32], medical image analysis [33,34], and machine health monitoring [37,38]. There are basically three important reasons why DL is becoming increasingly popular for researchers and extremely suitable for the fault diagnosis of REBs: 1) Data has become more easily accessible, allowing us to access a vast amount of industrial data more readily. One of the main reasons for this is that sensors are becoming more advanced and cost-effective, which enables their wider range of deployment [39,40]. 2) With the drastically increased computational ability of processing units, especially Graphics Processing Units and the significantly reduced cost of computing hardware, implementing, training and performing DL algorithms have become easier and quicker [41]. 3) Recent advances in ML algorithms and signal-processing technologies have led to great advance of DL algorithms and have resulted in successes of DL in various applications such as computer vision, phonetic recognition [42,43].

Deep learning algorithms seek to replicate the human brain's operation with interconnected layers of artificial neurons [44]. Data is input at the initial layer, passes through weighted connections and activation functions in hidden layers, and produces an output in the final layer [45]. These models adjust weights and biases through training and minimizing a loss function by comparing predictions to actual data. Two of the most representative types of DL structures are Autoencoders and Convolutional Neural Networks. Autoencoders are unsupervised learning models designed to capture essential features of input data by compressing it into a representation with lower dimension and then reconstructing it back to its original form. They are widely used for tasks like data compression and denoising [46]. Convolutional Neural Networks, or CNNs, are particularly suited for image and spatial data due to their hierarchical architecture that automatically extracts and learns features from data [47]. These two classes of algorithms have garnered significant research attention in recent years because of their versatility and effectiveness in a wide range of fields. Their capacity to handle large-scale and complex datasets has opened up new possibilities in the field of artificial intelligence. Therefore, in this paper, we will focus on summarizing and categorizing these two types of algorithms as well as their applications in the field of fault diagnosis.

This paper is constructed as follows. Section 2 introduces the models and application of Autoencoder, one of the most popular DL algorithms applied in bearing fault diagnosis. Section 3 describes the mechanism and applications of Convolutional Neural Network. Section 4 concludes the paper and discusses challenges as well as further explorations.

2. Autoencoder models and their applications

2.1 Autoencoder and Stacked Autoencoder

2.1.1 Autoencoder and Stacked Autoencoder algorithms

Autoencoder is a DL neural network structure which contains multiple layers that try to reconstruct the input signals. Basically, the layers can be classified into three types: input layers, output layers and hidden layers. The input layer is a vector constructed by a number of neurons based on the feature data abstracted from rotating machines. The vector is then encoded into the hidden layer which is a small representation of the data in the first layer. Finally, when forming the output layer, the hidden layer output is decoded into an output vector which contains the same number of neurons as the input layer. The reason why AE is implemented is that the data abstracted to reflect the condition of machines is usually high-dimensional which is time and effort consuming for computer to process, but Autoencoder is able to minimize the size of the data and help reduce computational complexity [29]. AE is mainly consisted of three steps: data preparation, encoding and decoding.

In the process of data preparation, a number of feature data is collected and turned into the input layer which can be considered as vector x. x is then mapped to a lower-dimensional representative in the next step, which results in the nonlinear hidden layer considered as vector y. The process can be described as follows:

$$y_e = f(W_e x + b_e)$$

Where f is a $sigmoid$ or $ReLU$ function.

The decoding step transfers vector y into the output layer vector z :

$$z = f(W_d y + b_d)$$

A loss function is also needed to train the AE network. It measures how well the Autoencoder is reconstructing the input data. In order to minimize the construction error and make the compressed layer accurate enough to represent the input data, the AE algorithm has to be trained by means of repeatedly applying the loss function. The function can be conveyed as the following equation:

$$L_{AE} = \|z - x\|^2$$

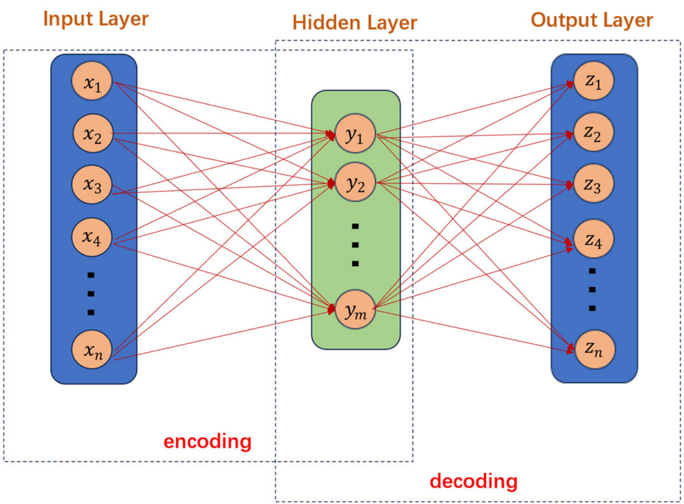

Fig. 1. Structure of Autoencoder

Stacked Autoencoder goes beyond a single Autoencoder. It consists of multiple layers of Autoencoders stacked on top of each other. Each layer learns a hierarchical representation of the input data. This stacking allows for the learning of increasingly abstract and complex features with the network deepens. The structure of SAE not only makes it more suitable for dimensionality reduction and feature extraction than single AE, but also makes it suitable for other functions such as hierarchical representation and unsupervised learning [35].

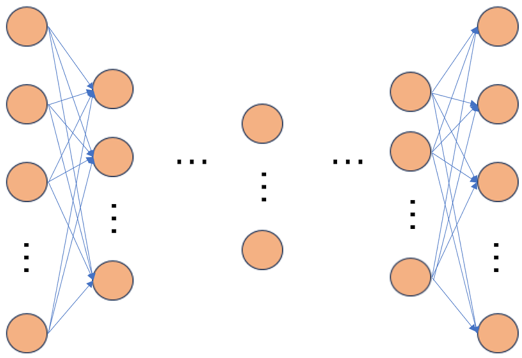

Fig. 2. Structure of Stacked Autoencoder

2.1.2 Autoencoder and Stacked Autoencoder applications

Autoencoders and Stacked Autoencoders offer versatile tools for feature extraction,

dimensionality reduction, noise robustness, anomaly detection, and transfer learning in fault diagnosis. Their ability to learn from unlabeled data and adapt to different domains makes them valuable assets in ensuring the reliability and safety of complex systems and machinery.

Mao et al. [50] introduced the Autoencoder Extreme Learning Machine to address the computational challenges of deep AE models. Autoencoder Extreme Learning Machine combines the efficiency of Extreme Learning Machine, a fast-training algorithm for single-hidden layer feed-forward neural networks, with high classification accuracy.

Jia et al. [51] tackled two traditional AE limitations: the extraction of similar features and the shift-variant nature of extracted features. They proposed a new structure termed NSAE-LCN which was based on Normalized Sparse Autoencoders (NSAE). The NSAE differs by using ReLU activation, eliminating bias, replacing KL divergence with norms, and adding a soft orthonormal constraint to the loss function.

In another research paper, Tao et al. [52] introduced the implement of a Stacked Autoencoder combined with softmax regression for the task of diagnosing bearing faults which showcases exceptional classification performance. Furthermore, this deep neural network exhibits robustness, effectively mitigating the impact of noise. Notably, the paper introduces an integrated deep neural network approach that encompasses ten distinct structural parameter networks, demonstrating strong generalization capabilities. Softmax regression, an extension of logistic regression to multiclass scenarios, offers a broader range of class labels compared to the binary choices in logistic regression. This attribute proves particularly valuable for resolving numerous complex multi-class classification challenges.

To address multimode fault diagnosis, a new deep AE model was introduced by Zhou et al. [53]. They first trained an SAE model for mode partitioning, followed by constructing multiple SAEs for observing data in each mode, determining the corresponding mode. At last, another SAE was built, taking the determined mode into account for bearing fault classification.

2.2 Denoising Autoencoder

2.2.1 Denoising Autoencoder algorithm

When implementing AE in practice, it is unavoidable that noise and variability exist in the data extracted in the input layer, which may cause the network to be inaccurate. This is why Denoising Autoencoder (DAE) is introduced. DAE is an extension of AE algorithm. The main structure of DAE is the same as the traditional AE network, but some improvement is added so that the DAE algorithm has more adaptive capacity to input data that may contain noise and corrupted data.

The improvement in a DAE algorithm is the way it constructs input layers. When it is processing an original dataset to form an input vector x, random noise is added to the input data. This noise can be in various forms, such as Gaussian noise, dropout, or other types of corruption. This results in the transformation of every origin input neuron x_i to a new corrupted neuron data $\overline{x_i}$. The Autoencoder is then trained to minimize the reconstruction error, which means to generate a clean version of the input data from the noisy version.

The noise injection process can be defined as follows:

$$x_c = x + \varepsilon$$

where ε denotes a random noise.

The new version of the loss function in a DAE algorithm can be described as:

$$L_{DAE} = \|z - x\|^2$$

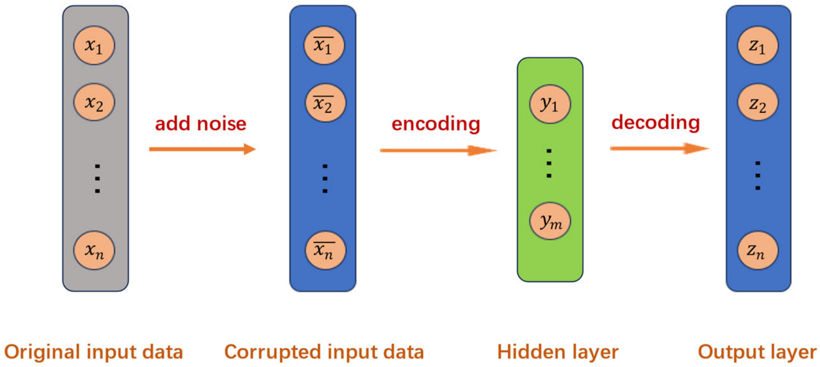

Fig. 3. Structure of Denoising Autoencoder

2.2.2 Denoising Autoencoder applications

Traditional AEs, including deep stacked models, face another challenge concerning the utility of the extracted features [54]. Xia et al. [55] proposed a method for automatic feature extraction using Stacked Denoising Autoencoders, incorporating masking noise to corrupt input data. Lu et al. [56] extended this approach by applying data corruption throughout all layers of the SAE.

In a publication, Shao et al. [57] presented an improved deep Autoencoder model that combines Denoising Autoencoders and Contrastive Autoencoders [58]. Initially, low-standard features are abstracted from raw vibration signals with a DAE, followed by the derivation of high-standard features through Contrastive Autoencoders.

2.3 Coupling Autoencoder

2.3.1 Coupling Autoencoder algorithm

A Coupling Autoencoder (CAE) is a neural network architecture that integrates the concepts of Autoencoders and coupling layers to capture complex data distributions and dependencies.

Usually, different sensors are applied at the same time which means that multiple types of signal datasets are abstracted together. A Coupling Autoencoder is therefore implemented to handle multisensory data. A CAE structure is also based on traditional utoencoder, but the hidden layer and decoding process are different.

First, two types of signals are abstracted by different sensors at the same time and they are transferred into input vector m and vector n. Then the two input layers are trained separately to form their own AE network. After that, a similarity measure is introduced to couple the two AE models. Consider the hidden layer of m is $\mathrm{h_m}$, while the hidden layer of n is $\mathrm{h_n}$, then the similarity measure can be described as follows:

$$S(n, m) = \|h_n - h_m\|^2$$

In order to train the joint representation of the two AE networks, a new loss function has to be applied:

$$\mathrm{L}_{\text{coupling}} = \alpha L_n + \beta L_m + \gamma S(n, m)$$

Where L_n is the AE loss function of input layer n, while L_m refers to the AE loss function of input layer m.

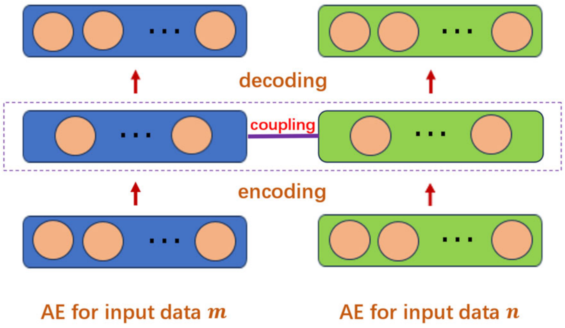

Fig. 4. Structure of Coupling Autoencoder

2.3.2 Coupling Autoencoder applications

In a study, Ma et al. [59] introduced a groundbreaking multimodal fusion deep neural network named Deep Coupling Autoencoder (DCAE), designed to uncover shared features between vibration and acoustic signals in the context of health status classification. This model applies a deep learning method based on CAE to merge multimodal signals gathered from diverse sensors. This integration of feature learning and multisensory data combination into a unified process is a key feature of this model. Additionally, the constructed deep architecture autonomously learns high-level features through a layer-by-layer training process, effectively capturing correlations between various signals. The results obtained from experiments on two multimodal datasets demonstrate the effectiveness of the utilization of DCAE in learning joint features from different modalities, leading to accurate health state classification compared to alternative approaches.

2.4 Sparse Autoencoder

2.4.1 Sparse Autoencoder algorithm

A Sparse Autoencoder is a type of neural network that uses sparsity as a means of compressing information. Compared with a traditional Autoencoder, a Sparse Autoencoder imposes sparsity constraints during the learning process to create an information bottleneck without reducing the number of neurons in the hidden layer. In a traditional AE model, some of the neurons in the hidden layer may not have significant effect on the network according to different input datasets. However, if these neurons are corrupted by some unexpected noises, the whole network and output can be affected greatly. Thus, it is desirable to implement a method so that only a few neurons in the encoder's hidden representation are activated, while most neurons remain inactive. This is why Sparse Autoencoders are applied. Sparsity constraint enables Sparse Autoencoders to learn more informative features because only a few neurons are responsible for capturing essential information from the input data.

In a Sparse Autoencoder, when a hidden layer which contains alpha nodes is formed, a sparsity constraint p is constructed based on the hidden nodes, which can be given as:

$$p = \frac{1}{\alpha} \sum_{i=1}^{\alpha} a_i$$

where a_i is one of the nodes in the hidden layer.

Then, in order to form the loss function in the training set and inactivate some nodes in the hidden representation, these nodes need to be set as close to zero as possible. Therefore, an additional penalty term is needed in the loss function

$$L_{sparse} = L + \beta \sum \mathrm{KL}(\tilde{p} || p)$$

$$\sum \mathrm{KL}(\tilde{p}||p) = \sum \tilde{p} \log\frac{\tilde{p}}{p} + (1-\tilde{p}) \log\frac{1-\tilde{p}}{1-p}$$

The sparsity parameter $\tilde{p}$ is obtained through dedicated methods and experiments, while the Kullback–Leibler divergence $\mathrm{KL}(\cdot)$ serves as a penalty metric, measuring the difference between the expected and observed distributions. If $\tilde{p} = p$, then $\mathrm{KL}(\tilde{p}||p) = 0$. As $\tilde{p}$ approaches 0 or 1, the KL-divergence tends towards infinity. The L refers to the original traditional AE loss function.

Sparse Autoencoders are commonly used in unsupervised learning tasks, such as feature learning, anomaly detection, and data compression. They are particularly useful when trying to discover underlying patterns or features in data while maintaining a compact representation.

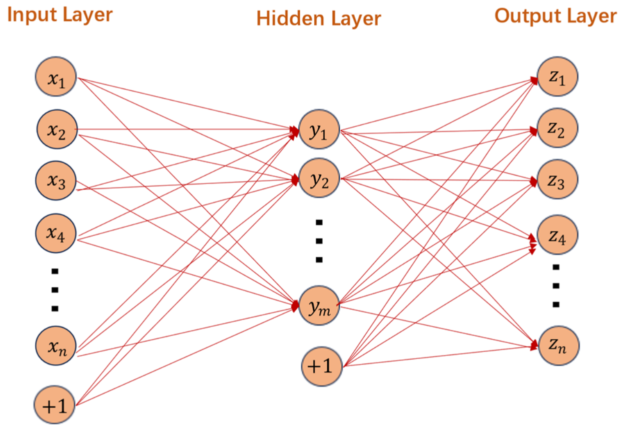

Fig. 5. Structure of Sparse Autoencoder

2.4.2 Sparse Autoencoder applications

Chen and Li [60] introduce a novel approach to bearing fault diagnosis using a multisensor feature fusion technique involving Sparse Autoencoder and DBN. The method involves amalgamating features derived from various accelerometers into a unified stream through unsupervised learning with Sparse Autoencoder. Subsequently, DBN is trained on these amalgamated features to perform classification. Extensive experiments were conducted, evaluating the method's performance on datasets. The results demonstrate that the proposed approach exhibits a superior recognition rate compared to other methods and displays lower sensitivity to training samples. Notably, in the case of dataset where vibration signals from different running speeds are intertwined, making it challenging to discern bearing status, this model still attains high classification accuracy and surpasses other feature fusion techniques. Moreover, the method holds promise for fusing data from diverse sensor types, suggesting practical applications that warrant further investigation.

The Winner-Take-All Autoencoder, presented by Makhzani and Frey [61], addresses sparse AE limitations. These limitations stem from the difficulty in selecting appropriate parameters for the Kullback-Leibler (KL) divergence-based sparsity penalty, as well as its incompatibility with rectified linear unit (ReLU) AEs. Winner-Take-All Autoencoder overcomes these issues by allowing flexibility in the target sparsity rate, ensuring efficient training and compatibility with various activation functions. Li et al. [62] modified Winner-Take-All Autoencoder by adding penalty terms to the hidden layer and weights.

3. Convolutional Neural Network and its applications

3.1 Convolutional Neural Network algorithms

Convolutional Neural Network (CNN) is a type of deep learning neural model built for tasks such as image recognition. CNNs are designed to replicate the workings of the mammalian visual cortex. These networks draw inspiration from the behavior of cells in the brain's visual cortex which contain cells that respond exclusively to local receptive fields [36,23]. Similar to the Autoencoder mentioned above, it also contains the processes of feature learning and feature classification. Basically, three types of layers are included in CNN: Convolutional Layer, Pooling Layer, and Fully Connected Layer. Convolutional Layer uses a set of small windows termed convolutional kernels to scan the feature datasets gained from sensors and obtain the feature maps of input data. The function of pooling layer is to minimize the dimension of feature maps and computational complexity. Fully Connected Layer is the layer responsible for feature classification. However, it shares the same classification step and way of operation as AE, which means it is the first two layers that makes CNN special.

In the Convolutional Layer, multiple learnable small filters termed convolutional kernels are used to slid across the input data and performing multiplication and summation. Each convolutional kernel has a set of weight parameters, which are learned through the training process. The operation can have a stride, which defines how much the kernel slides over the input data. A larger stride results in a reduction in the size of the output feature map, while a smaller stride maintains the size. The goal of the convolution operation is to capture features in the input data such as textures, shapes, and vibration by means of using these learned kernels. As a result of the convolution operation, the input data is transferred into a feature map. Each convolutional kernel generates one feature map. These feature maps capture information in the input data that matches the features defined in the kernel. By applying multiple kernels to the input data simultaneously, multiple features can be detected and multiple feature maps can be generated. CNNs typically accept multi-channel input (e.g., color images with three channels: red, green, blue) and produce multi-channel output. Each kernel performs a convolution on each channel and generates a feature map, and these feature maps are stacked along the channel dimension to form the multi-channel output.

In every layer of the network, if the input datasets have N matrices X_m which match the kernel implemented in the convolution process, the mechanism of the convolutional layer can be described as follows:

$$Y_i = f(\sum_{m=1}^{N} X_m * P_i + b_i)$$

Where Y stands for the output feature maps, while i is the index of the feature maps. P_i is the kernel used in each convolution process, and b_i is the bias matrix added to the convolutional outcome.

The Pooling Layer is used for reducing the spatial dimensions of the feature maps created by the Convolutional Layer while preserving the most representative data. This reduction helps decrease the computational load and the number of parameters in the network. The Pooling process achieve this goal by applying a small window called pooling window which has no learnable parameters. Pooling allows control over the output feature maps' size by specifying the pooling window's size and the length by which the window moves every time. Various of pooling methods can be implemented. Among them, two of the most popular methods are max pooling and average pooling. Max pooling retains the maximum value within each region, highlighting the most salient features, while the result of average pooling is the average value, which can provide a more smoothed representation of the data.

The pooling process can be described as follows:

$$Q_{max} = \max\left(Q_{input}: Q_{input} \in W\right)$$

$$Q_{average} = \frac{1}{n}\left(\sum_{Q_{input} \in W} Q_{input}\right)$$

Where the first equation describes max pooling, while the second equation describes average pooling. Q_{max} and $Q_{average}$ are the output of input data Q_{input} inside a pooling window, and W stands for the data included in the window.

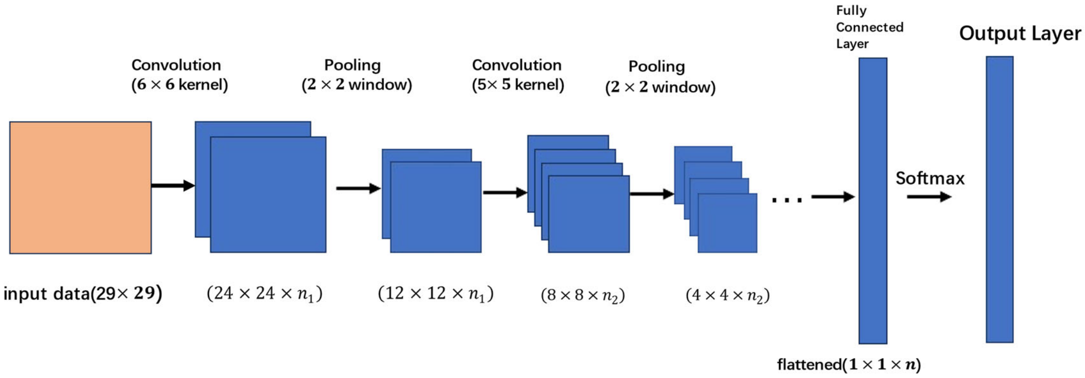

Fig. 6. Structure of Convolutional Neural Network

3.2 Convolutional Neural Network applications in bearing fault diagnosis

In recent times, the achievements of CNNs in diverse fields such as assignments associated with image recognition and classification have inspired their exploration in the realm of fault diagnosis. This is particularly evident in the context of rotating machinery, including components like bearings and gears. However, unlike 2-D image data, the input for these applications primarily comprises 1-D signals such as vibration and current data gathered from the machinery being monitored. To handle such 1-D input data, multiple solutions can be applied.

Jiang et al. [18] introduced a new architecture called Multiscale CNN designed for intelligent fault diagnosis in wind turbine gears under various operational circumstances. This network uses 1-D convolution to handle the 1-D signal data. The key innovation is the establishment of feature learning with multiscale, enabling autonomous and efficient acquisition of rich fault features from raw vibration signals. This approach outperforms traditional methods in feature learning, noise resilience, and classification. The Multiscale CNN system offers end-to-end fault diagnosis, particularly beneficial for large wind farms. It also provides a versatile framework for fault diagnosis in various machinery and industrial systems. In another study, Wen et al. [12] proposed another solution for 1-D input. A method transforming signal data to image data was implemented to divide the 1-D signal data into several sections and reconstruct it into the traditional 2-D image data.

In a paper, an CNN approach for fault diagnosis of rotating machinery with multiple sensor fusion was presented by Xia et al. [16]. Sensor fusion was accomplished in order to improve the accuracy and reliability of diagnosis by merging the raw signals gained from multiple sensors to construct the input data without the need for manual feature engineering. It employs techniques like mini-batch stochastic gradient descent and dropout during training to enhance efficiency and prevent overfitting, particularly when data is limited. Experimental studies on REBs and gears validate the method, demonstrating superior performance compared to traditional approaches. The approach allows for broad applications in fault diagnosis across different machine and fault types, even with limited prior knowledge and manually crafted features.

4. Conclusion

In this paper, a detailed review is presented focusing on the algorithms and applications of two deep learning network, Autoencoder and Convolutional Neural Networks, which have been widely used in the field of bearing fault diagnosis. Bearings play a crucial role in the functionality and efficiency of various mechanical systems, making their health and maintenance paramount. Traditional diagnostic methods based on ML, while effective to some extent, often require hand-craft feature extraction, specialized expertise, and human labor. The evolution of deep learning has revolutionized the realm of bearing fault diagnosis. For one thing, this research paper gives a clear review of Autoencoder, which is classified into four categories: Stacked Autoencoder, Denoising Autoencoder, Coupling Autoencoder, and Sparse Autoencoder. The applications of these

Autoencoder in fault diagnosis have also been illustrated. For another thing, the passage also provides a review on Convolutional Neural Networks by narrating its way of operating as well as its applications. In CNNs, the fusion of data from multiple sensors and the transformation of 1-D signal data into 2-D image data have expanded the horizons of fault diagnosis. At the meantime, there are still some challenges when dealing with fault diagnosis by means of DL networks. For example, in practical scenarios, a diverse range of bearing faults can emerge, often simultaneously. Hence, additional endeavors are necessary to validate the effectiveness of deep learning in diagnosing multiple bearing faults. As the industrial sector continues to evolve, the integration of deep learning networks in fault diagnosis will undoubtedly become more prevalent in the future, offering a more streamlined, efficient, and proactive approach to machinery maintenance and health monitoring.

References

[1] Ruyi, Huang., Jingyan, Xia., Bin, Zhang., Zhuyun, Chen., Weihua, Li. (2023). Compound Fault Diagnosis for Rotating Machinery: State-of-the-Art, Challenges, and Opportunities. Journal of Dynamics Monitoring and Diagnostics, doi: 10.37965/jdmd.2023.152

[2] R. Huang, J. Li, S. Wang, G. Li, and W. Li, “A Robust Weight-Shared Capsule Network for Intelligent Machinery Fault Diagnosis,” IEEE Trans. Ind. Informat., vol. 16, no. 10, pp. 6466–6475, Oct. 2020.

[3] B. Zhang et al., “Deep emulational semi-supervised knowledge probability imaging method for plate structural health monitoring using guided waves,” Eng. Comput., vol. 38, no. 5, pp. 4151–4166, 2022.

[4] C. Lanham, Understanding the Tests that are Recommended for Electric Motor Predictive Maintenance, Baker Instrument Company, 2002.

[5] S. Nandi, H.A. Toliyat, X. Li, Condition monitoring and fault diagnosis of electrical motors-a review, IEEE Trans. Energy Convers. 20 (4) (2005) 719–729.

[6] Sanjay, Kumar, Palei. (2023). <scp>Data-driven</scp> fault diagnosis approaches for industrial equipment: A review. Expert Systems, doi: 10.1111/exsy.13360

[7] (2023). Industrial Systems Fault's Diagnosis Approaches Adapted to the Context of the Smart Industry. doi: 10.1109/iraset57153.2023.10153043

[8] Rathnakar, Achary., Chetan, J., Shelke. (2023). Bearing Error Diagnosis Using Deep Learning and Convolution Neural Network. doi: 10.1109/INOCON57975.2023.10101200

[9] Yiwei, Cheng., Lin, Manxi., Jun, Wu., Haiping, Zhu., Xinyu, Shao. (2021). Intelligent fault diagnosis of rotating machinery based on continuous wavelet transform-local binary convolutional neural network. Knowledge Based Systems, doi: 10.1016/J.KNOSYS.2021.106796

[10] Gamal, Attiya., Ayman, El-sayed., A.T, El, Sayed., Ezz, El-Din, Hemdan. (2023). An Efficient Fault Diagnosis Framework for Digital Twins Using Optimized Machine Learning Models in Smart Industrial Control Systems. International Journal of Computational Intelligence Systems, doi: 10.1007/s44196-023-00241-6

[11] J. Zarei, M.A. Tajeddini, H.R. Karimi, Vibration analysis for bearing fault detection and classification using an intelligent filter, Mechatronics 24 (2) (2014) 151–157.

[12] Long, Wen., Xinyu, Li., Liang, Gao., Yuyan, Zhang. (2018). A New Convolutional Neural Network-Based Data-Driven Fault Diagnosis Method. IEEE Transactions on Industrial Electronics, doi: 10.1109/TIE.2017.2774777

[13] S. Singh, A. Kumar, N. Kumar, Motor current signature analysis for bearing fault detection in mechanical systems, Proced. Mater. Sci. 6 (2014) 171–177.

[14] Gerardo, Avalos., Sara, Fernández, Aguayo., Jose, de, Jesus, Rangel-Magdaleno., Mario, R., Arrieta, Paternina. (2022). Bearing fault detection in induction motors using digital Taylor-Fourier transform. doi: 10.1109/ICEM51905.2022.9910779

[15] F. Hemmati, W. Orfali, M.S. Gadala, Roller bearing acoustic signature extraction by wavelet packet transform, applications in fault detection and size estimation, Appl. Acoust. 104 (2016) 101–118.

[16] Min, Xia., Teng, Li., Lin, Xu., Lizhi, Liu., Clarence, W., de, Silva. (2018). Fault Diagnosis for Rotating Machinery Using Multiple Sensors and Convolutional Neural Networks. IEEE-ASME Transactions on Mechatronics, doi: 10.1109/TMECH.2017.2728371

[17] A. Malhi, R.X. Gao, PCA-based feature selection scheme for machine defect classification, IEEE Trans. Instrum. Meas. 53 (6) (2004) 1517–1525.

[18] Guoqian, Jiang., Haibo, He., Jun, Yan., Ping, Xie. (2019). Multiscale Convolutional Neural Networks for Fault Diagnosis of Wind Turbine Gearbox. IEEE Transactions on Industrial Electronics, doi: 10.1109/TIE.2018.2844805

[19] Hejun, Ye., Ping, Wu., Yifei, Huo., Xuemei, Wang., Yuchen, He., Xujie, Zhang., Jinfeng, Gao. (2022). Bearing Fault Diagnosis Based on Randomized Fisher Discriminant Analysis. Sensors, doi: 10.3390/s22218093

[20] P. Baraldi, F. Cannarile, F. Di Maio, E. Zio, Hierarchical k-nearest neighbours classification and binary differential evolution for fault diagnostics of automotive bearings operating under variable conditions, Eng. Appl. Artif. Intell. 56 (2016) 1–13.

[21] L.S. Dhamande, M.B. Chaudhari, Bearing fault diagnosis based on statistical feature extraction in time and frequency domain and neural network, Int. J. Vehicle Struct. Syst. 8 (4) (2016) 229–240.

[22] D.-S. Huang, A constructive approach for finding arbitrary roots of polynomials by neural networks, IEEE Trans. Neural Netw. 15 (2) (2004) 477–491.

[23] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[24] C. Li, R.-V. Sanchez, G. Zurita, M. Cerrada, D. Cabrera, R.E. Vásquez, Multimodal deep support vector classification with homologous features and its application to gearbox fault diagnosis, Neurocomputing 168 (2015) 119–127.

[25] F. Jia, Y. Lei, J. Lin, X. Zhou, N. Lu, Deep neural networks: a promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data, Mech. Syst. Signal Process. 72 (2016) 303–315.

[26] Md, Ifraham, Iqbal., Md., Saddam, Hossain, Mukta., Ahmed, Rafi, Hasan., Salekul, Islam. (2022). A Dynamic Weighted Tabular Method for Convolutional Neural Networks. IEEE Access, doi: 10.1109/ACCESS.2022.3231102

[27] Firas, Husham, Almukhtar., Asmaa, Abbas, Ajwad., Amna, Shibib, Kamil., Refed, A., Jaleel., Raya, Adil, Kamil., Sarah, Jalal, Mosa. (2022). Deep Learning Techniques for Pattern Recognition in EEG Audio Signal-Processing-Based Eye-Closed and Eye-Open Cases. Electronics, doi: 10.3390/electronics11234029

[28] Wookhee, Min., Randall, D., Spain., Jason, D., Saville., Bradford, W., Mott., Keith, Brawner., Joan, H., Johnston., James, C., Lester. (2021). Multidimensional Team Communication Modeling for Adaptive Team Training: A Hybrid Deep Learning and Graphical Modeling Framework.. doi: 10.1007/978-3-030-78292-4_24

[29] Franck, Galpin., Muhammet, Balcilar., Frédéric, Lefebvre., Fabien, Racape., Pierre, Hellier. (2023). Entropy Coding Improvement for Low-complexity Compressive Auto-encoders. doi: 10.1109/DCC55655.2023.00080

[30] Joel, Serey., Miguel, Alfaro., Guillermo, Fuertes., Manuel, Vargas., Claudia, Durán., Rodrigo, Ternero., Ricardo, Rivera., Jorge, Sabattin. (2023). Pattern Recognition and Deep Learning Technologies, Enablers of Industry 4.0, and Their Role in Engineering Research. Symmetry, doi: 10.3390/sym15020535

[31] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[32] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the AAAI, 2017, pp. 4278–4284.

[33] H. Greenspan, B. van Ginneken, R.M. Summers, Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique, IEEE Trans. Med. Imag. 35 (5) (2016) 1153–1159.

[34] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, I. Eric, C. Chang, Deep learning of feature representation with multiple instance learning for medical image analysis, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014, pp. 1626–1630.

[35] Chenliang, Liu., Yalin, Wang., Kai, Wang., Xiaofeng, Yuan. (2021). Deep learning with nonlocal and local structure preserving stacked autoencoder for soft sensor in industrial processes. Engineering Applications of Artificial Intelligence, doi: 10.1016/J.ENGAPPAI.2021.104341

[36] D. H. Hubel and T. N. Wiesel, “Receptive fields of single neurones in the cat’s striate cortex,” J. Phys., vol. 148, no. 3, pp. 574–591, Oct. 1959.

[37] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, R.X. Gao, Deep learning and its applications to machine health monitoring, Mech. Syst. Sig. Process. 115 (2019) 213–237.

[38] P. Tamilselvan, P. Wang, Failure diagnosis using deep belief learning based health state classification, Reliab. Eng. Syst. Saf. 115 (2013) 124–135.

[39] G., Scotti., Jamal, Saied-Walker., Noah, Marchal., Hang, Yu., Marjorie, Skubic. (2022). HTIDB: Hierarchical Time-Indexed Database for Efficient Storage and Access to Irregular Time-series Health Sensor Data. doi: 10.1109/EMBC48229.2022.9871855

[40] Tanja, Steinbach., Benjamin, Roehm., Reiner, Anderl. (2022). Development of a Guideline for the Sensor Layout Plan. doi: 10.54941/ahfe1001619

[41] Lalit, Kumar., Dushyant, Kumar, Singh. (2022). Hardware Response and Performance Analysis of Multicore Computing Systems for Deep Learning Algorithms. Cybernetics and Information Technologies, doi: 10.2478/cait-2022-0028

[42] Peng, Liang., Yu, Tang., Xiaoda, Zhang., Youhui, Bai., Teng, Su., Zhiquan, Lai., Linbo, Qiao., Dongsheng, Li. (2023). A Survey on Auto-Parallelism of Large-Scale Deep Learning Training. IEEE Transactions on Parallel and Distributed Systems, doi: 10.1109/TPDS.2023.3281931

[43] Carsten, Hartmann., Lorenz, Richter. (2023). Transgressing the boundaries: towards a rigorous understanding of deep learning and its (non-)robustness. arXiv.org, doi: 10.48550/arXiv.2307.02454

[44] Oscar, Montiel. (2023). Various Deep Learning Algorithms in Computational Intelligence. Axioms, doi: 10.3390/axioms12050495

[45] L. Deng, D. Yu, et al., Deep learning: methods and applications, Found. Trends Sig. Process. 7 (3–4) (2014) 197–387.

[46] Mingliang, Cui., Youqing, Wang., Xinshuang, Lin., Maiying, Zhong. (2021). Fault Diagnosis of Rolling Bearings Based on an Improved Stack Autoencoder and Support Vector Machine. IEEE Sensors Journal, doi: 10.1109/JSEN.2020.3030910

[47] Jeeraporn, Werapun., Anantaporn, Hanskunatai. (2023). Hierarchical Convolutional Neural Networks using CCP-3 Block Architecture for Apparel Image Classification. International Journal of Advanced Computer Science and Applications, doi: 10.14569/ijacsa.2023.0140624

[48] Liang, Ma., Zijing, Wu., Kaixiang, Peng. (2023). A Survey of Few-shot Learning-based Compound Fault Diagnosis Methods for Industrial Processes. doi: 10.1109/ICPS58381.2023.10128105

[49] Feng, Z., & Yang, S.. (2023). <i>A Semisupervised Method With Swarm Intelligence Optimization for Intelligent Fault Diagnosis</i>. <i>23</i>. https://doi.org/10.1109/JSEN.2023.3268131

[50] W. Mao, J. He, Y. Li, Y. Yan, Bearing fault diagnosis with auto-encoder extreme learning machine: a comparative study, Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci. 231 (8) (2017) 1560–1578.

[51] F. Jia, Y. Lei, L. Guo, J. Lin, S. Xing, A neural network constructed by deep learning technique and its application to intelligent fault diagnosis of machines, Neurocomputing 272 (2018) 619–628.

[52] S. Tao, T. Zhang, J. Yang, X. Wang and W. Lu, "Bearing fault diagnosis method based on stacked autoencoder and softmax regression," 2015 34th Chinese Control Conference (CCC), Hangzhou, China, 2015, pp. 6331-6335, doi: 10.1109/ChiCC.2015.7260634.

[53] F. Zhou, Y. Gao, C. Wen, A novel multimode fault classification method based on deep learning, J. Control Sci. Eng. 2017 (2017) 1–14.

[54] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion, J. Mach. Learn. Res. 11 (2010) 3371–3408.

[55] M. Xia, T. Li, L. Liu, L. Xu, C.W. de Silva, Intelligent fault diagnosis approach with unsupervised feature learning by stacked denoising autoencoder, IET Sci. Meas. Technol. 11 (2017) 687–695.

[56] C. Lu, Z.-Y. Wang, W.-L. Qin, J. Ma, Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification, Sig. Process. 130 (2017) 377–388.

[57] H. Shao, H. Jiang, F. Wang, H. Zhao, An enhancement deep feature fusion method for rotating machinery fault diagnosis, Knowl. Based Syst. 119 (2017) 200–220.

[58] S. Rifai, P. Vincent, X. Muller, X. Glorot, Y. Bengio, Contractive auto-encoders: explicit invariance during feature extraction, in: Proceedings of the Twenty-eighth International Conference on Machine Learning (ICML-11), 2011, pp. 833–840.

[59] Meng, Ma., Chuang, Sun., Xuefeng, Chen. (2018). Deep Coupling Autoencoder for Fault Diagnosis With Multimodal Sensory Data. IEEE Transactions on Industrial Informatics, doi: 10.1109/TII.2018.2793246

[60] Zhuyun, Chen., Weihua, Li. (2017). Multisensor Feature Fusion for Bearing Fault Diagnosis Using Sparse Autoencoder and Deep Belief Network. IEEE Transactions on Instrumentation and Measurement, doi: 10.1109/TIM.2017.2669947

[61] A. Makhzani, B.J. Frey, Winner-take-all autoencoders, in: Proceedings of the Advances in Neural Information Processing Systems, 2015, pp. 2791–2799.

[62] C. Li, W. Zhang, G. Peng, S. Liu, Bearing fault diagnosis using fully-connected winner-take-all autoencoder, IEEE Access 6 (2018) 6013–6115.