Review Article

Rongcai Wang, Enzhi Dong, Zhonghua Cheng, Zichang Liu*, and Xisheng Jia*

Transformer-based intelligent fault diagnosis methods of mechanical equipment: A survey

https://doi.org/10.1515/phys-2024-0015
received December 25, 2023; accepted March 22, 2024

Abstract: Transformer is extensively employed in natural language processing, and computer vision (CV), with the self-attention structure. Due to its outstanding long-range dependency modeling and parallel computing capability, some leading researchers have recently attempted to apply Transformer to intelligent fault diagnosis tasks for mechanical equipment, and have achieved remarkable results. Physical phenomena such as changes in vibration, sound, and heat play a crucial role in the research of mechanical equipment fault diagnosis, which directly reflects the operational status and potential faults of mechanical equipment. Currently, intelligent fault diagnosis of mechanical equipment based on monitoring signals such as vibration, sound, and temperature using Transformer-based models remains a popular research topic. While some review literature has explored the related principles and application scenarios of Transformer, there is still a lack of research on its application in intelligent fault diagnosis tasks for mechanical equipment. Therefore, this work begins by examining the current research status of fault diagnosis methods for mechanical equipment. This study first provides a brief overview of the development history of Transformer, outlines its basic structure and principles, and analyzes the characteristics and advantages of its model structure. Next it focuses on three model variants of Transformer that have generated a significant impact in the field of CV. Following that, the research progress and current challenges of Transformer-based intelligent fault diagnosis methods for mechanical equipment are discussed. Finally, the future development direction of Transformer in the field of mechanical equipment fault diagnosis is proposed.

*** Corresponding author: Zichang Liu,** Shijiazhuang Campus of Army Engineering University of PLA, 050003 Shijiazhuang, China, e-mail: zc_liu1997@aeu.edu.cn
*** Corresponding author: Xisheng Jia,** Shijiazhuang Campus of Army Engineering University of PLA, 050003 Shijiazhuang, China, e-mail: asd3v36@163.com
Rongcai Wang, **Enzhi Dong**, **Zhonghua Cheng:** Shijiazhuang Campus of Army Engineering University of PLA, 050003 Shijiazhuang, China

Keywords: Transformer, deep learning, mechanical equipment, computer vision, intelligent fault diagnosis

1 Introduction

With the continuous development of modern technology, the level of systematization, automation, and intelligence of mechanical equipment in industrial applications has been further improved. The functional structure has become increasingly diverse and complex, and is widely used in industries such as aerospace, transportation, power generation, automotive manufacturing, and machining, including specific applications such as aircraft engines, wind turbines, industrial gearboxes, high-speed trains, and construction machinery [1], *etc*. Due to the increasing demands for speed, load, and automation level of mechanical equipment in modern industrial production, equipment faults can easily lead to downtime, resulting in significant economic losses and even casualties. Statistics show that major accidents and economic losses caused by mechanical equipment faults account for approximately 38% of industrial production [2]. Therefore, real-time monitoring and fault diagnosis of mechanical equipment to ensure its normal operation and prevent serious accidents have become urgent issues in related industries. Meanwhile, the operation of mechanical equipment is often accompanied by complex physical phenomena, such as mechanical vibration, sound radiation, heat conduction, *etc*. These physical phenomena usually contain important information about the state of mechanical equipment, which is of great value to the research of intelligent fault diagnosis. Especially vibration signals, which usually contain dynamic response information inside mechanical equipment, can detect weak early faults. Accordingly, mechanical equipment fault diagnosis has become a critical part of system design and maintenance, with significant implications for improving economic efficiency. However, in industrial practice, modern mechanical equipment exhibits characteristics such as coupling, delay, and hierarchy in faults [3], thus requiring further exploration of fault diagnosis methods for mechanical equipment.

Fault diagnosis is the process of analyzing and determining the type and location of equipment faults. The general process of mechanical equipment fault diagnosis involves three steps: signal acquisition, feature extraction, and pattern recognition [4]. Traditional methods of mechanical equipment fault diagnosis primarily involve visualizing sensor data and establishing basic thresholds (such as temperature, vibration, and speed) to monitor the equipment's health status. However, these methods are unable to accurately and promptly identify early equipment failures. In recent years, the rapid advancement of artificial intelligence (AI) has led to increased attention and research from experts and scholars on intelligent fault diagnosis methods. These methods are based on traditional machine learning (ML), particularly deep learning (DL) [5]. Compared to traditional ML methods, DL networks have more complex structures and can automatically extract deep-level features from data, effectively distinguishing interference information and eliminating the influence of human factors. In the field of mechanical equipment fault diagnosis, DL models such as convolutional neural network (CNN) [6–8] and recurrent neural network (RNN) [9,10] have been extensively researched and applied, yielding fruitful results.

In engineering practice, CNN and RNN are commonly used to process image or sequence data. Fault diagnosis models based on CNN employ convolutional kernels to automatically extract fault features from the data, without being influenced by human factors. While CNN has strong image feature extraction capabilities, they tend to prioritize and depend on local feature information within the signal, often lacking explicit memory when processing sequence data. This results in an inability to effectively extract contextual features from sequence data, which in turn affects the accuracy of fault diagnosis.

Currently, mechanical equipment fault diagnosis that takes into account the time-related features of signals is primarily accomplished through the application of RNN and its enhanced variations. RNN can capture temporal information in sequential data and extract long-term dependencies due to their recurrent structure. They store information in their internal state and possess a certain memory capacity. However, RNN also has some limitations that require improvement, including high consumption of computing and storage resources, challenges in preserving information over long time intervals in sequences, inability to perform parallel computing, convergence difficulties, and gradient vanishing. Long short-term memory (LSTM) network is an enhanced version of RNN that tackles these issues by incorporating gate mechanisms and memory units, enabling it to effectively learn and retain long-term dependencies in sequential data. However, LSTM has a large number of model parameters, long training times, and requires a specific sequence for computation, which makes it unable to perform parallel computing. Additionally, it still has limitations in modeling large datasets.

In recent years, a network framework known as Transformer, which is based on the self-attention mechanism, has garnered significant attention from researchers [11]. Self-attention can encode and model the relationships between different positions in sequential data or images, enabling the model to efficiently extract key feature information from large data segments [12]. Transformer employs an internal self-attention mechanism to encode and model the positions within the sequence. It calculates results based on the similarity between different positions to obtain global contextual information, which is then input into the encoder and decoder layers for processing. This approach enables global feature extraction and long-range feature modeling. At the same time, Transformer eliminates all convolutional and recursive structures in the model, enabling efficient parallel computation of sequences. This makes it more effective in handling long-term, long-range dependencies and greatly improves the training efficiency of the model. Research works have shown that the performance of Transformer in tasks, such as natural language processing (NLP), computer vision (CV), and fault diagnosis, is comparable to or even surpasses that of various models based on CNN or RNN. It can be seen that, compared to commonly used DL methods, Transformer has the following four distinct advantages of global receptive field, modeling long-range dependencies, parallel computing capability, and big data processing capabilities. The comparison of the receptive fields of Transformer and CNN is depicted in Figure 1.

Figure 1 shows a comparison of the effective receptive fields between Transformer and CNN in the semantic segmentation task. SegFormer [13] and DeepLabv3+ [14] are typical models used for semantic segmentation tasks, with SegFormer being a Transformer-based model and DeepLabv3+ being a CNN-based model. These models demonstrate significant differences in the range and variation of effective receptive fields. The receptive field of a CNN is constructed through local connections and requires continuous iterations to gradually expand. In contrast, the effective receptive field of Transformer has the advantage of rapidly expanding its scope through global interaction mechanisms [15].

In summary, while some traditional ML and DL methods can effectively accomplish mechanical equipment fault diagnosis, there are still greater demands for accuracy and stability in fault diagnosis methods in real industrial settings. This is to minimize downtime losses and safety hazards resulting from mechanical equipment faults. In recent years, due to the

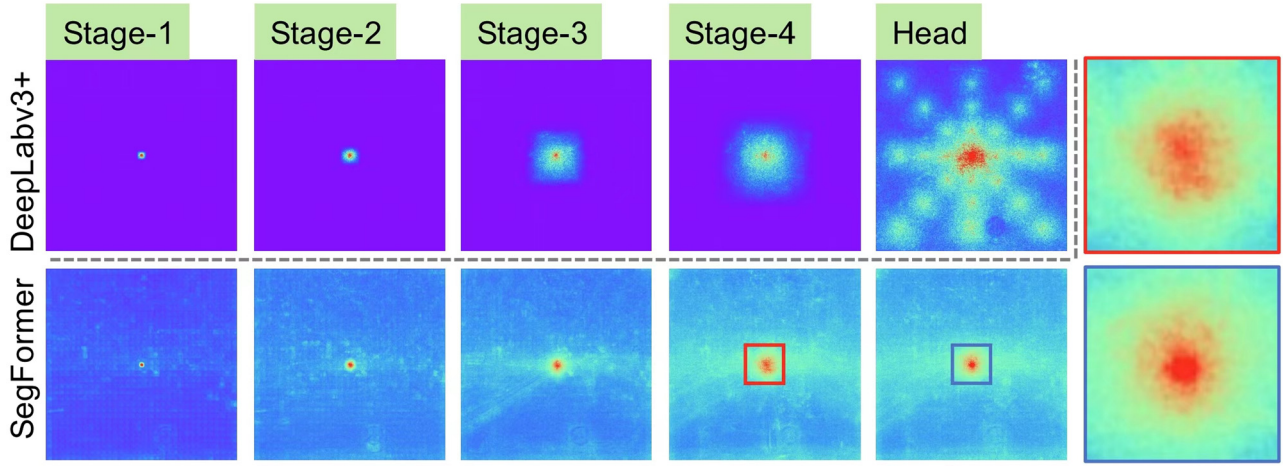

Figure 1: Comparison of effective receptive fields between Transformer and CNN in semantic segmentation tasks [13].

outstanding performance demonstrated by Transformer, an increasing number of studies have begun to incorporate Transformers into intelligent fault diagnosis for mechanical equipment to enhance the accuracy and robustness of fault diagnosis work.

At present, some research results have been achieved in Transformer-based intelligent fault diagnosis methods for mechanical equipment, but these studies are scattered in different literatures and lack systematic organization and generalization. Therefore, a comprehensive and systematic review of this emerging technology can help to understand the current status and development trend of its application in mechanical equipment fault diagnosis, promote the understanding and attention of academia and industry to this research field, and is of great significance in promoting the development and application of Transformer-based intelligent fault diagnosis methods for mechanical equipment.

The remaining part of this study is organized as follows: Section 2 reviews and analyzes the current research status and limitations of existing methods for diagnosing faults of mechanical equipment. In Section 3, the development history, basic structure, and principles of Transformer are first introduced. Then, three significantly effective model variants of Transformer in CV and commonly used public datasets for validating fault diagnosis methods are summarized. Finally, the research progress and application status of Transformer-based intelligent fault diagnosis methods for mechanical equipment are focused on and explored in detail. Section 4 provides a summary and outlook on the future development direction of Transformer models in the field of intelligent fault diagnosis for mechanical equipment. In Section 5, the study concludes with a summary of the entire content.

2 Progress and limitations of existing mechanical equipment fault diagnosis methods

Mechanical equipment fault diagnosis involves analyzing the equipment's performance to identify specific fault types. Common methods for mechanical equipment fault diagnosis typically include those based on physical models [16] and AI models [17]. The physical model-based method examines the evolution and alterations in fault mechanisms, such as wear, cracks, and fatigue. However, constructing the model necessitates expertise in fault mechanisms, professional domain knowledge, and a series of essential assumptions, making it challenging to accurately establish fault diagnosis models for complex systems. In recent years, the rapid advancement of computer technology and data science has led to increased attention on fault diagnosis methods based on AI models in the field of mechanical equipment failure diagnosis. These methods are gaining popularity due to their broad applicability, simplicity, and independence from the need to establish precise mathematical models like physical models. AI models employ a range of intelligent algorithms to analyze sensor data, extract feature information that accurately represents the equipment's state from extensive data, and subsequently identify the fault patterns of mechanical equipment [18]. Most AI models are typically designed and implemented using traditional ML methods. The latest generation of AI technologies, such as DL, has demonstrated significant advantages in feature extraction, knowledge acquisition, and intelligence. This has opened up new avenues for effective fault diagnosis of mechanical equipment, making it a popular research topic both domestically and internationally. This

study summarizes and analyzes the current research status of the three steps involved in the application of AI models for mechanical equipment fault diagnosis: signal acquisition, feature extraction, and pattern recognition.

In signal acquisition and feature extraction, the understanding and application of related physical principles, phenomena and models are very important. First, physical principles provide a theoretical basis for signal analysis and processing. For example, the wave and resonance principle can be used to analyze the frequency spectrum and time-frequency of vibration signals and extract the characteristic information related to faults. Second, the physical phenomenon is the direct manifestation of the mechanical equipment failure. Different fault types often correspond to different vibration phenomena, such as amplitude change, frequency deviation, and so on. Through the capture and analysis of these phenomena, it can provide a strong basis for fault diagnosis. Finally, physical models play a bridging role in intelligent fault diagnosis. By abstracting the actual mechanical device into a physical model, its operating state and fault process can be simulated, providing training data and verification criteria for DL models.

2.1 Signal acquisition

At present, in the process of mechanical equipment signal acquisition, the commonly employed signals are mainly vibration signals [19–21], sound signals [22], temperature signals [23], as well as oil analyzers [24], infrared thermography [25,26], *etc*. The sensitivity of each method in the process of mechanical equipment fault diagnosis and the corresponding maintenance costs [27] are shown in Figure 2.

As illustrated in Figure 2, it is evident that a high sensor sensitivity allows for the detection of certain equipment anomalies through signal analysis. However, it may not facilitate precise fault localization and elimination in practical mechanical equipment applications. As a result, the fault diagnosis accuracy is low, leading to increased maintenance costs. As sensor technology becomes more sensitive, it becomes easier to identify faulty parts and fault patterns in equipment, leading to reduced maintenance costs. However, the likelihood of equipment fault also increases. The process of acquiring and analyzing signals forms the basis for identifying subsequent fault patterns. Various sensor technologies can be compared and analyzed based on their advantages and disadvantages, as illustrated in Table 1.

Among the aforementioned sensor technologies, vibration signals are most widely applied in mechanical equipment fault diagnosis due to their ease of measurement and the fact that they contain important dynamic information about the mechanical equipment, such as the reciprocating motion of piston-connecting-rod assemblies, crankshafts, gear rotations, and so on [28].

2.2 Feature extraction

Feature extraction is a crucial step in the fault diagnosis process, which helps to identify patterns and structures in the data and provide better inputs for outputting fault

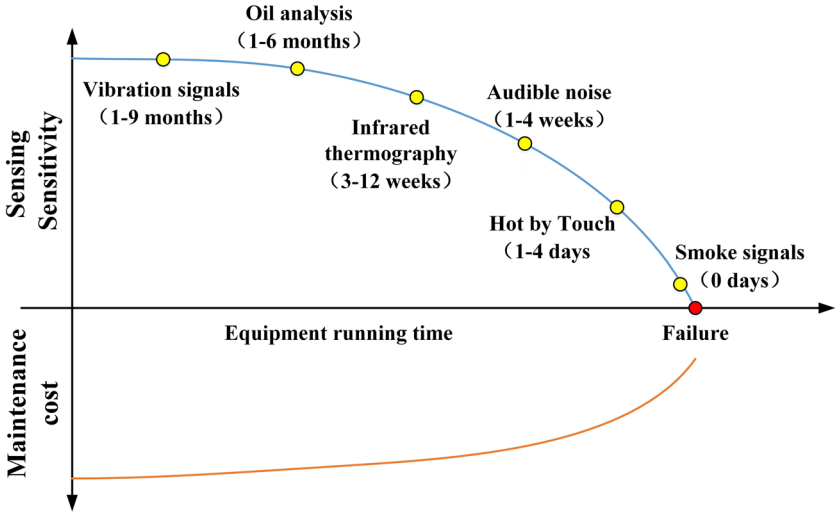

Figure 2: Sensitivity and corresponding maintenance cost of different sensor technologies in mechanical fault diagnosis.

types. The quality of feature extraction directly impacts the effectiveness of fault diagnosis. Therefore, in terms of signal feature extraction, a large number of theoretical studies and application have been explored. At present, the commonly employed feature extraction methods are mainly as follows:

1) **Time domain features.** Time domain features refer to the characteristics and statistics of a signal in the time domain. They describe the instantaneous changes and temporal information of the signal, effectively characterizing the distribution and concentration trend of the data. Common time domain features include mean [29], peak [30], root mean square (RMS) [31], entropy [32], and others.
2) **Frequency domain features.** Frequency domain features refer to the properties of a signal in the frequency domain, which are used to characterize the frequency content and distribution of the signal. This process yields the frequency domain signal, from which statistical features can be extracted. Commonly deployed frequency domain statistical feature parameters include the center of the frequency [33], frequency variance [34], and RMS frequency [35], among others. Other commonly applied frequency domain analysis methods include power spectrum [36], cepstrum [37], and other analysis techniques, which have yielded some results in the field of fault diagnosis.
3) **Time-frequency domain analysis.** The vibration signals and sound signals of mechanical equipment typically exhibit non-smoothness and nonlinearity. This characteristic makes it unsatisfactory to rely solely on statistical parameters in the time and frequency domains for fault diagnosis. Therefore, the feature extraction method based on time-frequency domain analysis is introduced into mechanical equipment fault diagnosis. Commonly employed methods for time-frequency domain analysis include the short-time Fourier transform (STFT) [38], wavelet transform (WT) [39], instantaneous frequency (IF) [40], and Hilbert–Huang transform [41].

   These methods can decompose a one-dimensional signal into different components in the time and frequency domains, creating two-dimensional time-frequency maps, and extracting the corresponding features for pattern recognition. In time-frequency domain analysis, the commonly used feature parameters include energy features, spectral features, phase features, modulation features, and IF features. Furthermore, DL methods can be employed to automatically extract the high-level abstract features from the time-frequency diagram. These features can then be directly input into the pattern recognition module for fault diagnosis.
4) **Image features.** With the rapid development of CV technology, mechanical equipment fault diagnosis methods based on image processing have been emphasized by more and more scholars [4]. The commonly used methods for extracting image features mainly fall into the following four types:

   Color feature extraction, including color histogram, color moments, color mean-variance, *etc.*; texture feature extraction, including grayscale covariance matrix, local binary pattern, Gabor filter, *etc.*; shape feature extraction, including edge detection, contour extraction, shape descriptor *etc.*; DL-based image feature extraction, which mainly employs deep neural network models to learn advanced feature representations from images, including CNN, RNN, stacked auto-encoder (SAE), deep belief network, and so on.
5) **Text features.** Text features are some meaningful information extracted from the text to represent the characteristics and content of the text. The following are some common text features and their extraction methods:

Text length: the number of characters or words in the text is regarded as a feature; Syntactic features: features based on syntactic structure, such as dependency relationships, syntactic trees, *etc.*; Topic models: used to identify the main themes or topics in the text, common methods include latent Dirichlet allocation [42] and non-negative matrix factorization [43]; Text sentiment features: used to determine the emotional tendencies of the text, such as emotional vocabulary, emoticons, *etc.*; DL features: for text data, first, bag-of-words (BoW), TF-IDF, Word2Vec, and GloVe, and other word vector models are employed to convert the text into vector representation [44–47], then more abstract and advanced representations are extracted by some DL models.

2.3 Pattern recognition

After signal acquisition and feature extraction, it is finally necessary to carry out fault pattern recognition and output fault diagnosis results. Fault pattern recognition is commonly known as fault classification, which presents the fault diagnosis results in the most intuitive form to management decision makers. Different pattern recognition methods are suitable for different types of data and have different classification capabilities. Therefore, for the pattern recognition problem in mechanical equipment fault diagnosis, the majority of researchers have made a lot of attempts and explored different pattern recognition methods, which

Table 1: Summary and analysis of commonly used sensor technologies

Sensor types	Advantages	Disadvantages
Vibration signal sensor	1) Easy to operate and measure 2) The signal contains important dynamic information of mechanical equipment	1) Contact detection requires mechanical equipment to be shut down when arranging sensors, resulting in certain shutdown losses 2) Some mechanical equipment works in complex and poor environments, such as high temperature and high pressure, which can easily lead to sensor measurement distortion or even damage
Sound signal sensor	1) Non-contact and non-destructive testing 2) Easy to operate, flexible installation position	1) During the signal acquisition process, there will be complex noise pollution, making it difficult to separate and extract fault signals 2) The transmission path is complex, and the representation of signal features lacks a unified standard, which is not conducive to improving detection accuracy 3) Unable to work in a vacuum environment
Temperature signal sensor	1) Wide temperature measurement range, simple structure, and signal transmission over long distances 2) Small size, good long-term stability, and high accuracy	1) Local measurement cannot fully characterize the state changes of mechanical equipment 2) For moving objects, small targets, or objects with very small thermal capacity, significant measurement errors may occur
Oil analyzer	1) Stable and reliable, with high analysis accuracy, capable of monitoring early wear related faults in equipment 2) It can directly detect the performance indicators and pollution status of lubricating oil and grease	1) More engineering practical experience and professional knowledge are required, and the fault standards of mechanical equipment are difficult to define 2) The application is relatively limited, and if the fault of mechanical equipment is not caused by wear and tear, it is often difficult to determine its health status
Infrared thermal camera	1) Non-contact and non-destructive testing 2) Flexible installation position, capable of remote monitoring over a wide range 3) Wide temperature measurement range, visualized equipment status, and easy to understand	1) The parameter settings are relatively complex, and it is difficult to accurately measure the surface emissivity of mechanical equipment 2) Industrial grade infrared thermal camera is expensive and has high signal acquisition costs

are mainly divided into two categories: traditional ML methods and DL methods, as shown in Table 2.

2.4 Limitation analysis

To summarize, different pattern recognition methods have different characteristics, and in practical applications, it is necessary to choose appropriate pattern recognition methods according to different environments and data types. In Section 2.3, the advantages and disadvantages of commonly employed traditional ML methods and DL methods and their applicable environments are summarized, and the following conclusions can be drawn:

1) Traditional ML methods require manual selection and extraction of features, a process that usually requires domain knowledge and experience. The selection of features may affect the performance of the model, and extracting features manually requires a lot of time and effort;
2) Traditional ML methods face challenges when dealing with high-dimensional data; as the dimensionality of the features increases, the computational complexity and storage requirements of the model increase, and it is also prone to the dimensionality catastrophe problem;
3) Traditional ML methods usually assume that there is a linear or near-linear relationship between features and labels, which may not fit the model well for data with a nonlinear structure or where there are complex relationships between features, leading to performance degradation;
4) DL methods such as CNN have problems such as limited convolutional kernel sensing field, difficulty in capturing image global information and gradient disappearance with the increase in network layers; RNN has limitations such as inability to compute in parallel

Table 2: Summary and analysis of commonly used pattern recognition methods

Methods	Advantages	Disadvantages
K-nearest neighbor (KNN) [48,49]	1) The principle of KNN is simple and easy to understand and implement 2) KNN is a parameter-free learning algorithm, which does not require an explicit training process, and only needs to save the training data, and can directly use the sample data for classification or regression prediction 3) KNN does not need to make any premise assumptions about the distribution of the data, has a wide range of applications, and works well for multi-categorization problems	1) KNN requires the calculation of the distance from each sample point to the sample to be classified, so the amount of calculation is large and is not applicable to large-scale datasets 2) Requiring storage of all sample data, high space complexity, and larger storage space 3) Only applicable to the case where the data type is numerical or categorical, not applicable to other types of data, such as text, images, *etc*
Decision tree (DT) [50–52]	1) The results of DT can be presented graphically, which is easy to understand 2) DT can automatically select the most important features without the need for manual feature selection, and performs well when dealing with high dimensional data 3) Relative to other algorithms, DT requires less preprocessing of data and can handle data containing missing values or discrete features	1) DT is prone to overfitting on training sets, leading to poor performance on new data 2) DT is insensitive to feature correlation and may fail to capture complex data relationships 3) DT can only handle classification problems with one output variable, cannot solve multi-output problems, and is not applicable to large-scale datasets
Random forest (RF) [53–55]	1) RF is an integrated learning method that consists of multiple DTs, and the results are obtained by voting or averaging, which can effectively reduce the risk of overfitting of individual DT and improve the accuracy of the model 2) RF has good robustness to missing data and outliers and can handle various types of data 3) DT in RF can be generated in parallel, which can fully utilize the computational resources and improve the training speed	1) RF needs to generate multiple DTs, and each DT needs to perform feature selection and node division, so it requires high computational resources 2) RF is less interpretable, and it can provide the ranking of feature importance, but cannot provide the specific relationship between features 3) RF is not suitable for processing datasets with imbalanced categories, because each DT is constructed based on independent random samples without regard to category imbalance
Hidden Markov model (HMM) [56,57]	1) HMM can effectively deal with time series data, such as speech recognition and NLP 2) HMM can capture potential patterns and hidden states in data, thus enabling modeling and analysis of complex systems 3) The parameters of HMM include state transfer probabilities, observation probabilities, *etc*., which are highly interpretable	1) The core assumption of HMM is the Markov property that the current state depends only on the previous state, an assumption that may not hold in some real-world scenarios, resulting in a limited model fitting ability 2) HMM assumes that observation sequences are generated from potential states, and thus there is some randomness in the prediction process 3) HMM is sensitive to the size of the state space, and when the state space is large, the parameter estimation and computational complexity increase significantly
Support vector machine (SVM) [58,59]	1) SVM can effectively deal with high-dimensional feature space and nonlinear problems by introducing kernel function to map input data to high-dimensional feature space 2) SVM uses only a small number of support vectors to determine the decision boundary, does not rely on the whole dataset, which reduces the computational complexity and is suitable for few-shot datasets 3) SVM has a solid mathematical theoretical foundation and is highly interpretable	1) SVM requires a great deal of experimentation and experience in tuning various parameters, such as kernel functions, penalty parameters, and regularization parameters 2) SVM has limited ability to process large-scale datasets and may encounter problems of insufficient memory and long computation time 3) SVM was originally designed for binary classification problems and is not effective in dealing with multi-categorization problems
Naive Bayes (NB) [60]	1) For large-scale datasets, due to the simple probability-based model, the NB algorithm has high computational efficiency 2) Due to its probability-based model, the results of NB have good interpretability	1) NB assumes that features are independent of each other, which is often untenable in practical problems and may lead to degraded classification performance

(Continued)

Table 2: *Continued*

Methods	Advantages	Disadvantages
	3) NB has better robustness to missing data	2) NB needs to determine the prior probability, so unreasonable assumptions about the prior model can lead to poor classification performance 3) NB is not flexible enough for the selection of input features to deal with situations where there are complex relationships between features
DL [61–68]	1) DL can perform feature selection and extraction automatically without human involvement, which allows it to handle very complex tasks and large-scale datasets 2) DL can learn higher-level feature representations through complex nonlinear transformations between layers, capturing more abstract and complex features in the data 3) DL is highly scalable and can improve performance by adding more neural network layers and parameters	1) DL typically requires large amounts of labeled data for training and has high requirements on the quality and diversity of the data 2) DL requires a large amount of computational resource for training and inference, especially when dealing with large datasets and complex models 3) The decision-making process of DL is usually black-box and it is difficult to explain how a model arrives at a certain prediction, limiting the use of DL in certain domains, such as healthcare and law, where explanatory applications are required

and difficulty in capturing dependencies in sequences at long distances; SAE has deficiencies such as difficulty in training, lack of theoretical support, *etc*.; DBN has deficiencies such as susceptibility to local optimal solutions, difficulty in parameter tuning, and risk of overfitting; and the above problems have greatly affected the effectiveness of the DL methods in fault diagnosis.

3 Application of Transformer in mechanical equipment fault diagnosis

Transformer [11] was first applied to machine translation tasks in NLP with remarkable results. In recent years, with the advancement of research, Transformer has also been innovatively introduced into the CV field, such as image enhancement, image generation, image classification, object detection, and image segmentation tasks [69–73], which created a new milestone in the CV field, and the development history of Transformer and the key model are shown in Figure 3.

By analyzing Figure 3, it can be seen that in June 2017, Vaswani *et al.* [11] proposed a Transformer framework based only on the self-attention mechanism for the first time, which demonstrated excellent performance in the field of NLP. In February 2018, Parmar *et al.* [74] proposed the Image Transformer model, which was an application of Transformer to the CV domain the first time. Since then, the visual Transformer model has been rapidly developed and many landmark results have emerged. For example, in May 2020, Carion *et al.* [72] constructed Detection Transformer (DETR), a new end-to-end object detection framework, and for the first time, Transformer was used for

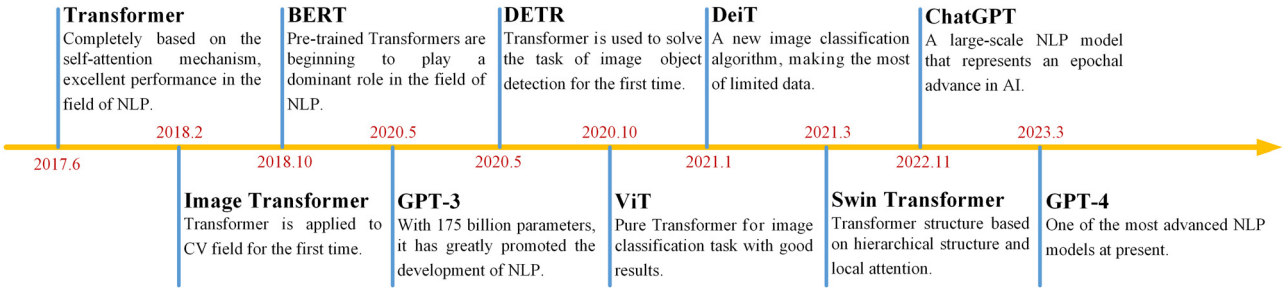

Figure 3: The evolution and key models of Transformer.

solving the image object detection task. In October 2020, Dosovitskiy *et al.* [75] proposed an image classification model, namely, Vision Transformer (ViT), based on the Transformer framework. By introducing a self-attention mechanism, ViT completely abandons the traditional convolutional operation. In January 2021, Facebook AI and the Sorbonne University [76] jointly developed a new efficient image classification algorithm Data-efficient image Transformers (DeiT). DeiT can perform adaptive feature extraction for each sample and thus is not affected by differences in the number of samples. This enables DeiT to perform well in small-sample learning tasks and to efficiently utilize limited data for classification. In practice, Transformer-based image classification models are complex in structure, have many parameters, and are computationally expensive; therefore, in March 2021, Liu *et al.* [77] proposed a new image classification model, Swin Transformer. It was inspired by the success of CNN in the image domain in its design, and prior knowledge from CNN was applied to the Transformer, including localization, multi-scale, and hierarchical design, and achieved optimal results on several image classification and object detection tasks, which was awarded as the best paper of ICCV2021, showing the great potential of Transformer.

3.1 Network structure and fundamentals of Transformer

Transformer is a sequence-to-sequence DL model based on the self-attention mechanism. Before Transformer was proposed, RNN was the most widely used model in the field of NLP [78], and its structure is illustrated in Figure 4.

As indicated in Figure 4, RNN is trained sequentially, and its structure contains loop units, and the output at a certain moment comes from the memory of the previous generations of loops (hidden state) and the current input state, *i.e.*, the loop units can memorize the previous information and input it into the next loop unit, so RNN is able to extract correlations between contextual features. However, there are three obvious flaws in the design of RNN:

1) RNN can only perform computation sequentially in order, which weakens the parallel computing ability of the model.
2) Long-term memory in the hidden state tends to weaken or even be lost with iterations, so it is difficult for RNN to establish long-term dependencies between distant features (*e.g.*, words from different parts of an utterance).
3) The training time of the model increases as the length of the input sequence becomes longer. This is because during training, the model needs to process each input token sequentially and use the information from the previous token when generating the next one. Therefore, longer input sequences require more computation and time to process.

Meanwhile, Transformer solves the above problems by introducing a self-attention mechanism. Transformer consists of three main modules such as encoder, decoder, and positional encoding [11]. As shown in Figure 5, the encoder generates the input encoding and the decoder receives all the encodings and uses them to merge the contextual information to generate the output sequence. Each module of Transformer is described in detail here.

3.1.1 Encoder-decoder

Transformer employs an encoder-decoder model architecture that avoids loops, as shown in Figure 6. The first part is the encoder, which consists of six identical encoder layers stacked together. Each encoder layer consists of two sublayers, namely, the multi-head self-attention (MSA) and the feed-forward neural network (FFN). MSA can focus on different positions in the input sequence, capturing global contextual information. FFN is used to perform nonlinear transformations on the features at each position. The entire encoder gradually extracts abstract representations of the input sequence through the stacking of multiple encoder layers.

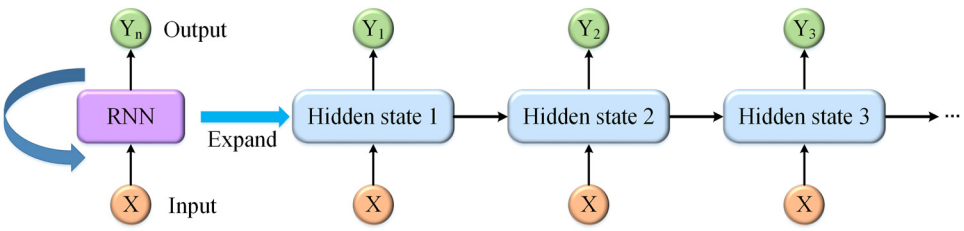

Figure 4: The conventional framework RNN.

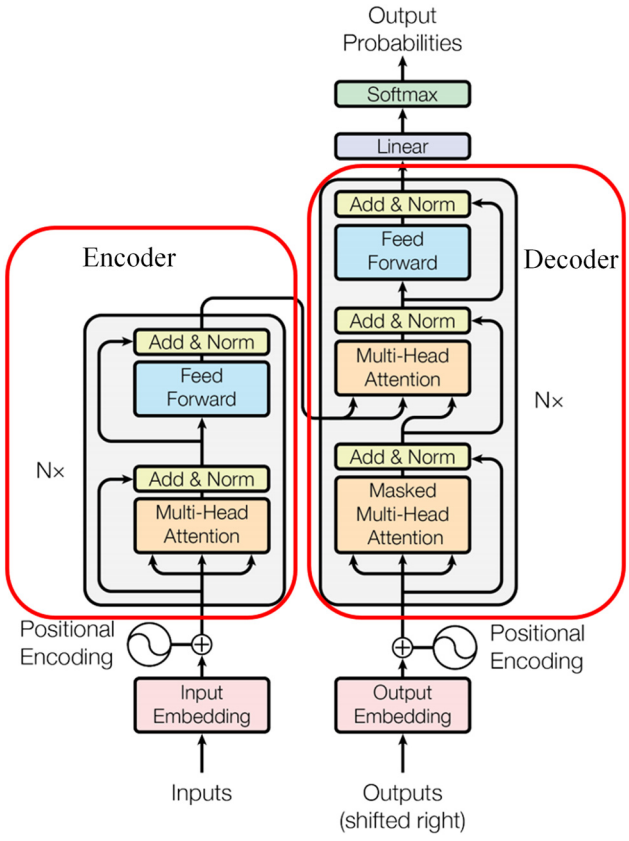

Figure 5: The original structure of Transformer [11].

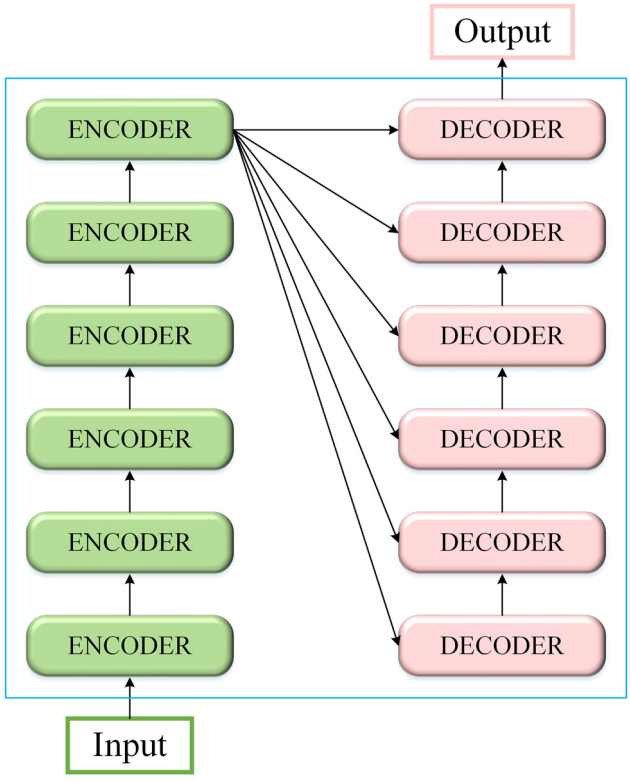

Figure 6: The structure of encoder-decoder.

Next is the decoder. After the input data are processed by the encoder layers, they are passed to each layer of the decoder to compute attention scores. The decoder also consists of six identical decoder layers, each of which comprises three sub-layers: MSA, masked multi-head attention (MMSA), and FFN. MSA retrieves information about the input sequence from the output of encoders, helping the decoder generate the correct output sequence. MMSA is primarily employed to mask or hide information at certain positions to prevent the model from overly relying on previous tokens during generation, thus improving the model's generalization ability and suppressing information leakage.

3.1.2 Feed-forward network

FFN is a fully connected feed-forward neural network that is added after the self-attention layer of each encoder and decoder. It receives the output of the self-attention layer as input and then outputs a new representation vector containing higher level semantic information. The computational procedure of FFN is as follows:

$$\mathrm{FFN}(\boldsymbol{X}) = \boldsymbol{W}_2\sigma(\boldsymbol{W}_1\boldsymbol{X}), \tag{1}$$

where $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are the linear transformation matrices of the first and second fully connected layers, respectively, σ denotes the nonlinear activation function, and the dimension of the hidden layer is d_h = 2,048.

In FFN, a two-layer fully connected structure is used and a ReLU activation function is employed between the two layers. Specifically, in each FFN, the input representation vector is first linearly transformed through one fully connected layer, then nonlinearly transformed through a ReLU activation function, and finally linearly transformed through another fully connected layer to obtain the output. The advantage of FFN is that they can extract higher-level semantic features of the inputs through the nonlinear transformations of the multiple layers, which improves the expressive power of the models. In addition, since the computation of FFN is performed independently, the training process of the model can be accelerated by parallelization.

3.1.3 Position encoding

Since Transformer does not contain any recursive and convolutional structures to capture the positional information

of different words in the text, some relative or absolute positional information about the tokens in the sequence has to be added in order for the model to be able to record the sequential relationships between the sequence data. To this end, position encoding is added to the bottom of the stack of the encoder and decoder, numbering the position of each word in the text, with each number corresponding to a word vector, respectively, and by combining the position vectors and the word vectors, certain positional information is added to each word. Compared with the sequential input method of RNN, Transformer can input data in parallel and retain the positional relationship between the data, which improves the computation speed and reduces the storage space requirement. Moreover, the dimensions of the position encoding and the input sequence embedding vectors are the same, so both can be summed up. Currently, there are various methods for position encoding [79], Transformer is realized by using sine and cosine functions with different frequencies [11], which can preserve the relative relationship of position information, and the specific operation can be expressed as follows:

$$\mathrm{PE}_{(\mathrm{pos},2i)} = \sin(\mathrm{pos}/10{,}000^{2i/d_m}), \tag{2}$$

$$\mathrm{PE}_{(\mathrm{pos},2i+1)} = \cos(\mathrm{pos}/10{,}000^{2i/d_m}), \tag{3}$$

where pos denotes the position of the word in the text each time, i represents the dimension, d_m is the dimension of the position encoding, $2i$ is the even dimension of the position encoding, and $2i + 1$ represents the odd dimension of the position encoding (i.e. $2i \le d, 2i + 1 \le d$). It follows that each dimension of the position encoding corresponds to a sinusoidal wave with a geometric progression of wavelengths from 2π to $10{,}000{-}2\pi$.

3.1.4 Self-attention

Attention mechanism [80] can be traced back to research in the field of neuroscience. Through the attention mechanism, the human brain is able to selectively focus on specific information and filter out irrelevant information when faced with a complex external environment. This mechanism allows us to better process and understand the information we receive. In the field of AI, the attention mechanism is introduced into DL models to improve the performance of the models. By introducing the attention mechanism, the model can dynamically assign different attention weights according to different parts of the input data and selectively process the important information. This mechanism can help the model better understand the input data and extract the key information in it. In recent years, the attention mechanism has been widely employed in speech recognition [81], machine translation [82], and image processing [83].

Self-attention [84] improves the attention mechanism in some details:

1) **Objects of attention:** the self-attention mechanism focuses on the correlation between different parts within the input information, while the attention mechanism mainly focuses on the correlation between the elements within the input utterance.
2) **Information processing method:** the information processing method of the self-attention mechanism does not rely on external information, but mainly on the characteristics of the input information itself. The attention mechanism, on the other hand, can utilize external information, such as contextual information, in its processing.
3) **Performance:** the self-attention mechanism can better capture the internal relevance of the input information, which is especially suitable for solving the long-distance dependency problem, whereas the traditional attention mechanism may encounter difficulties in dealing with this kind of problem.

By introducing the self-attention mechanism, Transformer completely abandons convolutional and recursive operations in its structure and relies only on the self-attention mechanism for global feature information extraction. The original Transformer proposes the scaled dot-product attention (SDPA) [11]. The basic structure of SDPA is depicted in Figure 7.

As is shown in Figure 7, Let $Y \in \mathbb{R}^{n\times d_m}$ be a sequence $(y_1, y_2, y_3, \ldots, y_n)$ containing n elements, where d_m is the dimension of the element m. In the self-attention mechanism, three trainable weight matrices are defined as query matrix $\mathbf{W}^Q \in \mathbb{R}^{n\times d_q}$, key matrix $\mathbf{W}^K \in \mathbb{R}^{n\times d_k}$, and value matrix $\mathbf{W}^V \in \mathbb{R}^{n\times d_v}$. Each element of the input sequence $Y \in \mathbb{R}^{n\times d_m}$ is linearly projected to each of the three weight matrices to generate three new vectors: the query vector (Query, $\boldsymbol{Q}$), the key vector (Key, $\boldsymbol{K}$), and the value vector (Value, $\boldsymbol{V}$), which are computed as follows:

$$Q = \boldsymbol{Y}\boldsymbol{W}^Q,\ \mathrm{K} = \boldsymbol{Y}\boldsymbol{W}^K,\ \mathrm{V} = \boldsymbol{Y}\boldsymbol{W}^V, \tag{4}$$

SDPA, on the other hand, computes the dot product of the query vector $\boldsymbol{Q}$ and the key vector $\boldsymbol{K}$ with scaling, then performs Softmax normalization, and finally multiplies it with the value vector $\boldsymbol{V}$ to obtain the output matrix. The specific calculation formula is as follows:

$$\mathrm{Attention}\,(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\cdot\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}, \tag{5}$$

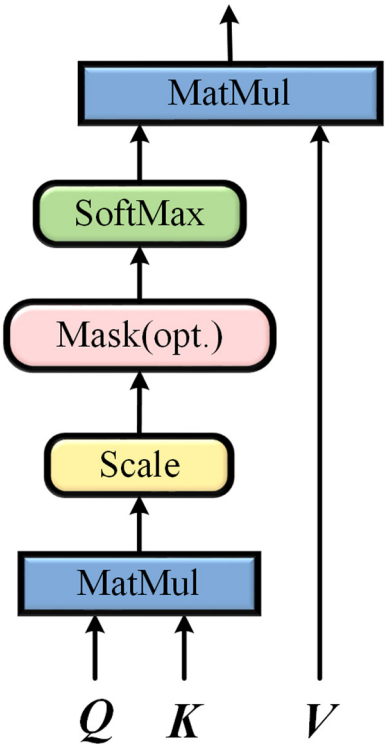

Figure 7: The basic structure of SDPA.

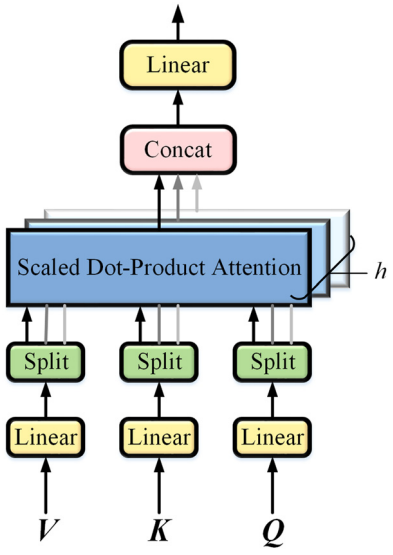

Figure 8: The basic structure of MSA.

where d_k represents the dimensions of the vectors $\boldsymbol{Q}$ and $\boldsymbol{K}$, and $\sqrt{d_k}$ is the scale factor (square root of the dimension of the key vector $\boldsymbol{K}$).

Each word in the self-attention mechanism computes an attention score (weight) that reflects the word's relevance to the other words in the sequence, thus capturing global information about the sequence and preserving information about long-term dependencies between different words.

MSA is an extension of self-attention and consists of multiple independent self-attention layers (heads), each of which has independently trainable weight matrices $\boldsymbol{W}_i^Q$, W_i^K, and $\boldsymbol{W}_i^V$. The basic structure of MSA is shown in Figure 8.

The essence of MSA is to obtain multiple sets of queries, keys, and values by mapping the input sequences into different subspaces by linear transformations while ensuring that the number of parameters is overall constant. The specific calculation process is as follows:

$$\boldsymbol{Q}_i = Y\boldsymbol{W}_i^Q, \boldsymbol{K}_i = Y\boldsymbol{W}_i^K, \boldsymbol{V}_i = Y\boldsymbol{W}_i^V, \tag{6}$$

$$\boldsymbol{Z}_i = \mathrm{Attention}(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i), \quad i = 1, 2, 3, \ldots, H, \tag{7}$$

$$\mathrm{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{Concat}(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{Z}_3, \ldots, \boldsymbol{Z}_H)\boldsymbol{W}^P, \tag{8}$$

where h is the number of heads of MSA, $\boldsymbol{Z}_i$ represents the output vector of each self-attention head, $\boldsymbol{W}^P$ denotes the linear transformation matrix of the output vector of each self-attention head. $\boldsymbol{Q}_i$, $\boldsymbol{K}_i$, and $\boldsymbol{V}_i$ can be considered as multiple splits of $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ in the self-attention performed in different subspaces.

3.2 Transformer-based image classification models

In the CV field, the original Transformer is not commonly used, because the original Transformer is mainly applied for sequential data processing tasks. With the deepening of the research, researchers have continuously improved the Transformer and gradually applied it to image processing tasks, with good results. According to the different design focuses and application areas of the model, this study summarizes the classical models of Transformer in the field of image processing and their applications, as shown in Table 3.

The Transformer model has demonstrated strong ability in processing sequence data (such as texts), and image data can also be regarded as a kind of sequence data, only that this sequence is two-dimensional. Inspired by this, some researchers have gradually applied the Transformer model to the field of image processing and developed many visual Transformer models with excellent performance. Among them, image classification is one of the most common applications of visual Transformer models. Image classification is the process of distinguishing images belonging to different categories and determining the category labels according to

Table 3: Summary and analysis of Transformer's classical models in image processing field

Task type	Classical models	Design focus
Image classification	ViT [75]	Image chunking, embedding, and serialization
	DeiT [76]	Knowledge distillation, self-supervised learning
	Swin Transformer [77]	Nested structures, local attention
	TNT [85]	Pyramidal layer-by-layer feature extraction
	PVT [86]	Nested structures, local attention
	T2T-ViT [87]	Pyramidal layer-by-layer feature extraction
	DeepViT [88]	Local prior, tokens-to-token mechanism
	CaiT [89]	Convolutional mapping, re-attention mechanism
	CrossViT [90]	Multi-scale feature fusion and cross-attention mechanism
Object detection	DETR [72]	New positional coding methods, teacher-student strategy
	TSP [91]	Ensemble-based global objective function, dichotomous matching
	ACT [92]	Pure encoder architecture, match distillation
	FPT [93]	Adaptive feature clustering, locally sensitive hashing
Image segmentation	SETR [94]	Multi-scale feature fusion, up-sampling strategy
	Segmenter [95]	Point-by-point linear decoder, adaptation to different resolutions
	SegFormer [13]	Hierarchical feature representation, lightweight full MLP decoder

the meaning and contextual information embedded in the images, which is an important fundamental work for other image processing tasks such as object detection and image segmentation.

To improve the processing efficiency of the visual Transformer model, researchers have made a series of improvements on the basis of the original Transformer, and in this study, some of the visual Transformer models with excellent performance are selected to be briefly introduced and analyzed. At present, typical Transformer-based image classification models mainly include ViT and its variants. This study focuses on introducing and summarizing the research progress of ViT and its variants.

3.2.1 Vision Transformer

ViT is the first successful application of Transformer in the field of image classification, which exceeded the state-of-the-art CNN models ResNet [96] and EfficientNet [97] at that time in terms of image classification performance, and the model architecture is shown in Figure 9.

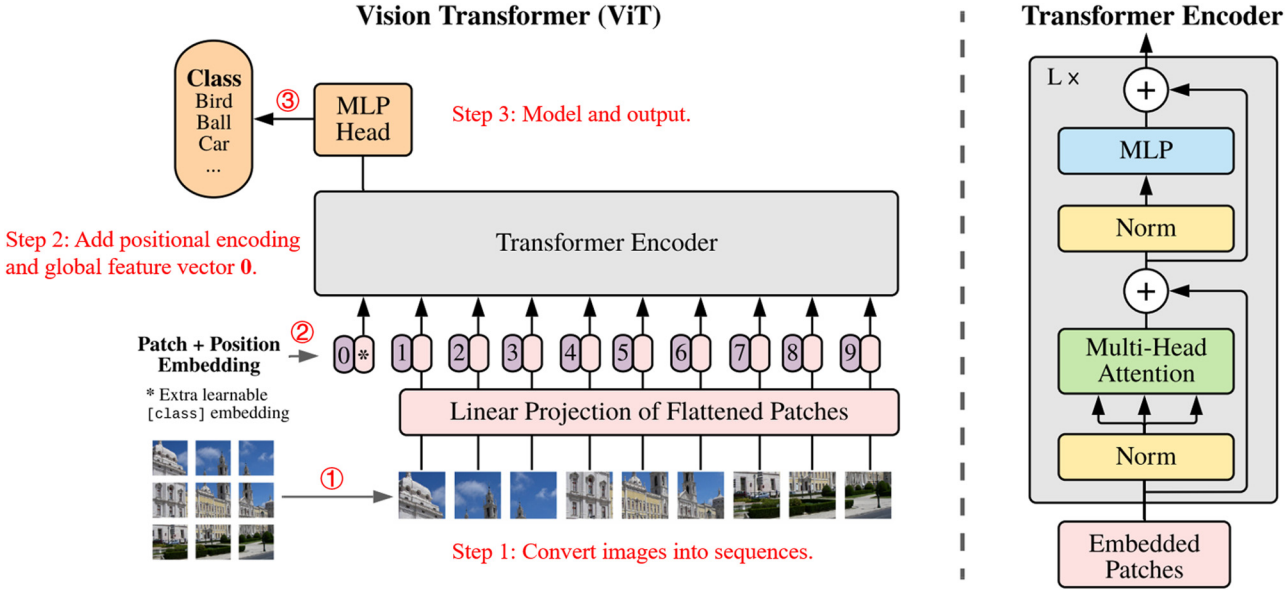

Figure 9: The model structure of ViT [75].

The key to applying ViT for image classification is to convert the image into sequential data. ViT segments the input image into a series of patches, each containing a portion of the image information. Each patch is then transformed into a vector representation, which is called embedding.

3.2.2 Data-efficient image Transformers

Based on knowledge distillation and self-supervised learning, DeiT needs less data and arithmetic power to achieve image classification results on the ImageNet dataset that are comparable to top CNN models. The model structure of DeiT is shown in Figure 10.

As shown in Figure 10, the overall training process of DeiT consists of four stages: data preparation, model architecture, self-supervised pre-training, and supervised fine-tuning. Compared with the traditional ViT, DeiT can achieve better results with fewer samples or even no samples.

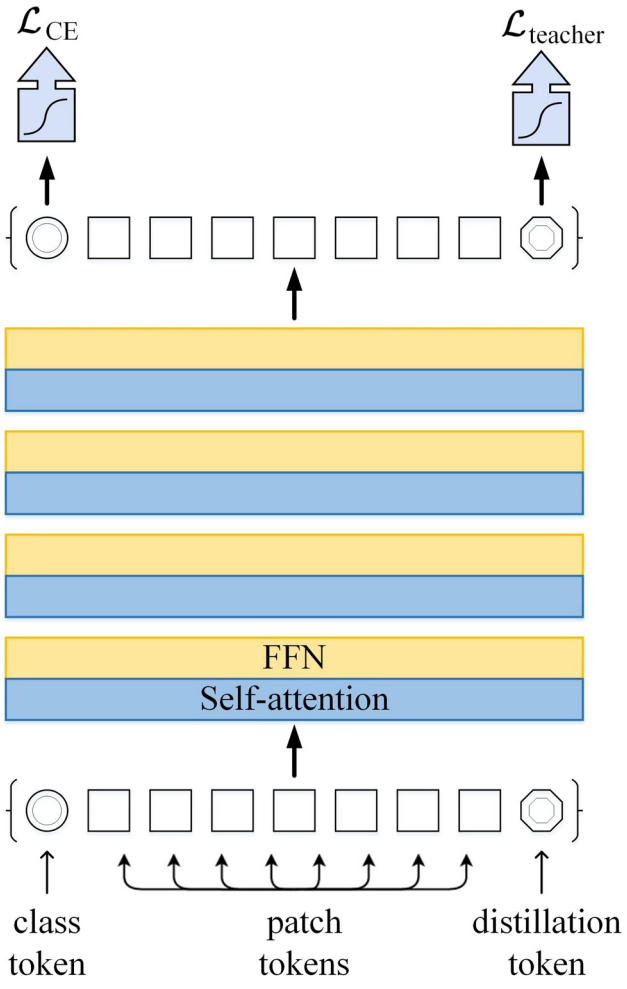

Figure 10: The model structure of DeiT [76].

3.2.3 Swin Transformer

Traditional ViT models usually need to divide the whole image into multiple fixed-size image patches and then perform global attention computation. However, this approach tends to destroy the local information of the image. Swin Transformer introduces a hierarchical Transformer structure, as shown in Figure 11, which realizes multi-scale feature extraction by dividing the input image into a series of non-overlapping windows, and then the local self-attention computation within the windows is performed.

As illustrated in Figure 11, the computational complexity of Swin Transformer varies linearly with the size of the input image, whereas the computational complexity of the ViT model varies as a square multiple of the size of the input image [98].

3.2.4 Tokens-to-token vision Transformer (T2T-ViT)

The core idea of T2T-ViT is to progressively refine image features through a recursive token transformation mechanism to improve the classification performance. The model structure of T2T-ViT is shown in Figure 12.

T2T-ViT improves the traditional ViT model by introducing a recursive token conversion mechanism and a deep narrow network structure, which improves the performance and efficiency of image classification.

3.2.5 Cross-attention multi-scale vision Transformer (CrossViT)

CrossViT improves the traditional ViT by combining multi-scale feature fusion and cross-attention mechanism. The model structure of CrossViT is shown in Figure 13.

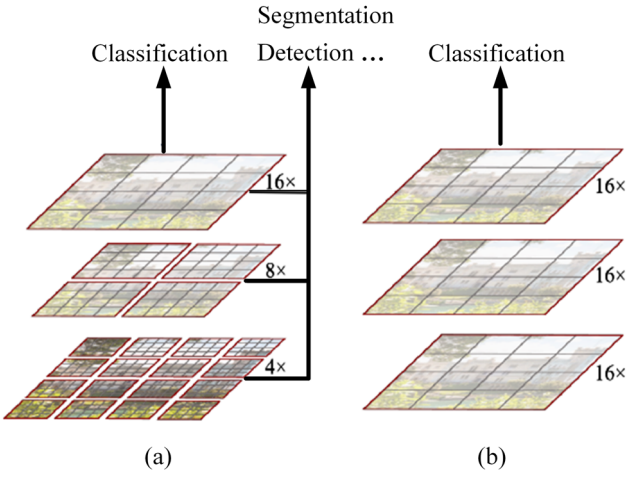

Figure 11: The hierarchical structure of Swin Transformer [77]. (a) Swin Transformer. (b) ViT.

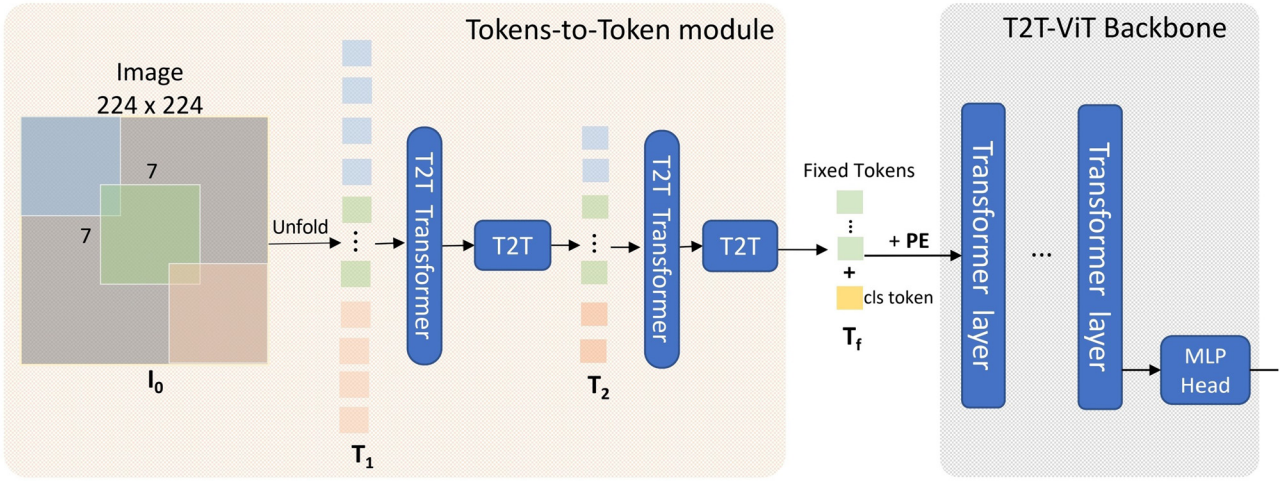

Figure 12: The model structure of T2T-ViT [87].

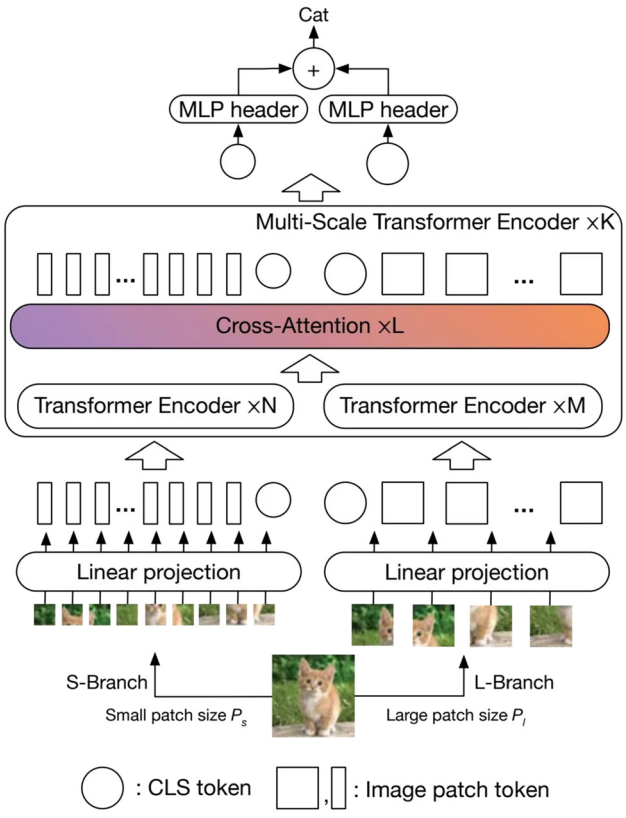

Figure 13: The model structure of CrossViT [90].

As shown in Figure 13, CrossViT introduces a two-branch ViT structure, where each branch handles image features at different scales. This mechanism allows information to be exchanged between two branches, thus enabling feature fusion. CrossViT can make full use of both local and global information of an image to improve the classification performance.

3.2.6 Pyramid vision Transformer (PVT)

PVT combines the pyramid structure in CNN and the Transformer's self-attention mechanism. The core idea of PVT is to construct a feature pyramid structure with different resolutions in the Transformer. The model structure of PVT is shown in Figure 14.

As demonstrated in Figure 14, PVT employs a down-sampling operation similar to that in CNN to gradually reduce the resolution of the feature maps, thus constructing feature pyramids with different resolutions.

3.3 Transformer-based intelligent fault diagnosis methods of mechanical equipment

The original Transformer model is not directly applicable to processing image data, as it was originally designed for NLP tasks where the input to the model is a set of word vector matrices. However, in recent years researchers have proposed a number of improvements and extensions to the Transformer model that make it usable for visual tasks such as image recognition.

By summarizing the relevant literature, it can be found that there are usually two processing ideas when using Transformer-based methods for fault diagnosis of mechanical equipment. One is to pre-process the input one-dimensional fault signals (vibration, sound, *etc.*) to convert the original one-dimensional signals into a form suitable for Transformer input, which is convenient for feature extraction. Since the vibration, sound, and other signals of mechanical equipment are generated by the

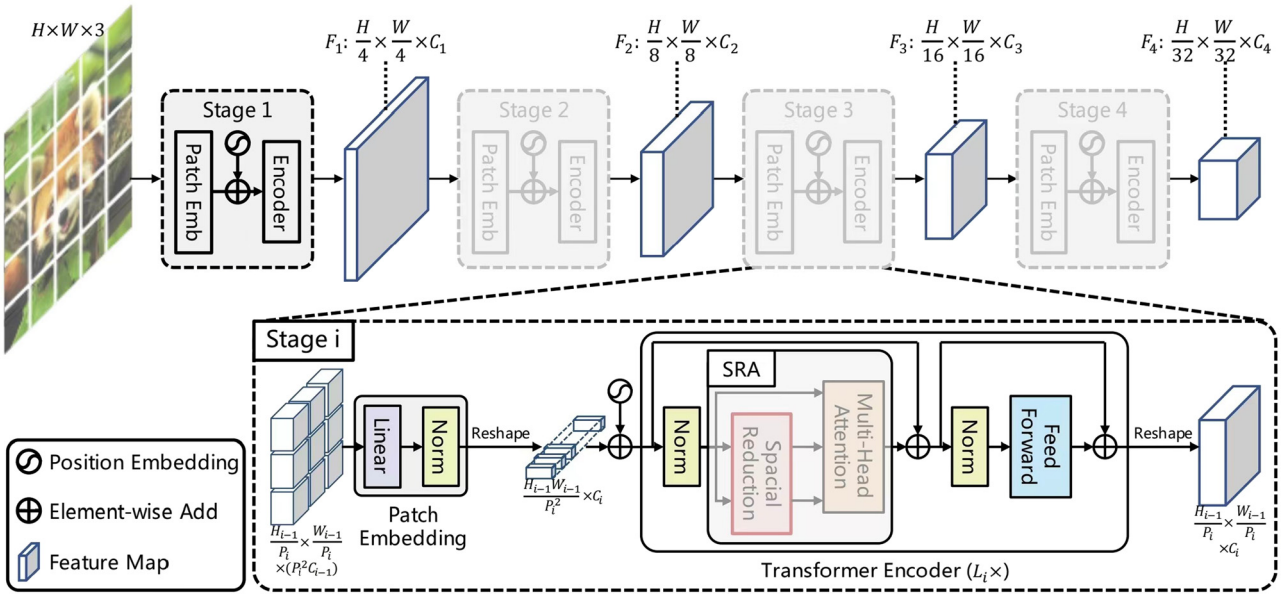

Figure 14: The model structure of PVT [86].

motion and dynamics characteristics of mechanical equipment, and their features include frequency, amplitude, phase, waveform, *etc*., they can be analyzed and processed by sensors and mathematical methods; whereas, natural language involves vocabulary, sentence structure, grammatical rules, and contexts, *etc*., and its understanding and generation need to rely on the cognitive and linguistic capabilities of the human, so the original Transformer cannot directly carry out mechanical equipment fault diagnosis by processing vibration, sound, and other signals. Second, the original one-dimensional signal is converted into a two-dimensional image by employing the relevant time-frequency conversion method, and then the two-dimensional image data are input into visual Transformer models such as ViT and Swin Transformer for training and fault pattern recognition.

3.3.1 Verification of fault diagnosis methods

Currently, the datasets used in the validation session of Transformer-based mechanical equipment fault diagnosis methods mainly include four kinds of public datasets, such as the Case Western Reserve University (CWRU) bearing dataset [99], XJTU-SY rolling bearing accelerated life test dataset [100], University of Connecticut (UCONN) gearbox fault dataset [101], and Harbin Institute of Technology (HIT) aero-engine inter-shaft bearing fault dataset [102], which contain a variety of bearing fault patterns, rolling bearing full-life cycle vibration data, multiple gearbox fault data, and the vibration signal of rotors and casings, respectively. The arrangement of the data acquisition equipment corresponding to the above four datasets is shown in Figure 15.

The CWRU bearing dataset contains a total of ten prefabricated faults of rolling bearings, which contain the normal condition as a special fault pattern, and has been widely applied in fault diagnosis research for rotating machinery. The XJTU-SY dataset is a widely employed rolling bearing accelerated life test dataset, which is collected and compiled by a team from Xi'an Jiaotong University in cooperation with Zhejiang Changxing Shengyang Science and Technology Co. Different from the CWRU bearing dataset, this dataset records the full life cycle vibration data of 15 rolling bearings under three operating conditions. The UCONN dataset is a gearbox dataset collected and arranged by the University of Connecticut from a two-stage gear box. The UCONN dataset contains 936 samples and 9 failure modes, which are normal condition, missing teeth, root fracture, contact surface spalling, and five kinds of tooth tip defects in different degrees. The HIT dataset, including three states of inter-shaft bearings, is proposed based on a real aero-engine by replacing the inter-shaft bearing with artificial fault, driven by motors and equipped with a lubricating system.

Through in-depth analysis, it can be concluded that the above public datasets reveal the following physical phenomena and physical principles:

1) **Physical phenomena.** In the normal operation of mechanical equipment, due to the periodic movement of rotating

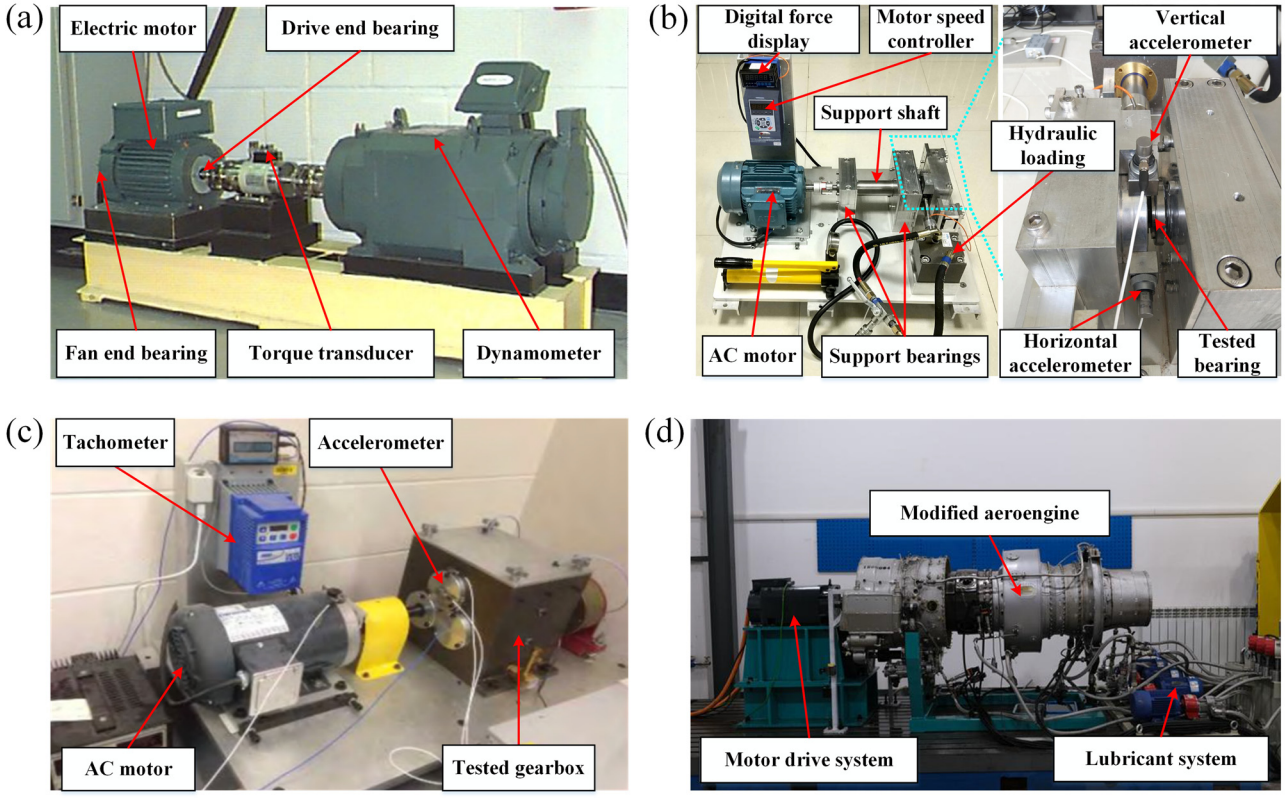

Figure 15: Data acquisition equipment: (a) CWRU dataset; (b) XJTU-SY dataset; (c) UCONN dataset; and (d) HIT dataset.

parts (such as bearings, gears, *etc.*), periodic vibration signals will be generated. These signals usually appear as distinct peaks on the spectrogram, related to the speed of the mechanical equipment and the natural frequencies of the components. When a component in the mechanical equipment fails (such as bearing damage, gear teeth broken, *etc.*), a transient impact component is generated in the vibration signal. These impact components often have a wide frequency band and high energy, which is an important basis for fault diagnosis. With the deterioration of mechanical equipment performance and the development of faults, the amplitude and frequency components of vibration signals tend to change. For example, the amplitude may gradually increase, while certain frequency components may disappear or new frequency components may appear.

2) **Physical principles.** When the vibration frequency of a mechanical device is close to the natural frequency of a component, a resonance phenomenon will occur, resulting in a significant increase in vibration amplitude. This phenomenon is of great significance in fault diagnosis because the resonance frequency is often related to the type of fault. In the vibration process of mechanical equipment, energy is transferred and converted between the various components. When a component fails, it will lead to abnormal energy transmission and conversion, which reflects the fault information in the vibration signal. The vibration signal of mechanical equipment can be regarded as the propagation of mechanical waves in the medium. The wave principle can be used to explain the attenuation, scattering, and interference of vibration signals in the process of propagation, and provide theoretical support for fault diagnosis.

The performance of different methods on these four public datasets are summarized in Table 4.

Analysis of Table 4 shows that the fault diagnosis accuracy of the HIT dataset is significantly lower than that of the other three datasets, which is due to the fact that the HIT dataset is closer to the actual fault diagnosis situation of mechanical equipment, which is more challenging compared to the dataset obtained under laboratory conditions, and the dataset provides a new benchmark for the validation of mechanical equipment fault diagnosis methods.

3.3.2 Related literature review of Transformer-based intelligent fault diagnosis methods

Driven by the first Transformer-based intelligent fault diagnosis research idea of mechanical equipment, Jin *et al.* [106]

Table 4: Performance of different methods on these 4 public datasets

Methods	Classification accuracy (%)
CWRU dataset	
BPNN [103]	81.35
DBN [104]	88.20
1DCNN [105]	97.32
TST [106]	98.63
Diagnosisformer [107]	99.85
SiT [108]	99.46
TAR [109]	99.90
SViT [110]	97.56
XJTU-SY dataset	
NKH-KELM [111]	95.56
DCN [112]	99.31
AlexNet [113]	99.58
LSTM [113]	98.65
CWT-2DCNN [114]	99.40
TST [106]	99.78
UCONN dataset	
AE [113]	95.13
DAE [113]	93.76
BPNN [113]	95.13
LSTM [113]	88.74
ResNet18 [113]	85.84
TST [106]	99.51
HIT dataset	
CNN [106]	83.13
LSTM [106]	85.41
TST [106]	71.07

proposed a fault diagnosis method for rotating machinery based on time series Transformer (TST) to solve the problem of long-term dependence of traditional CNN and RNN-based fault diagnosis models. The overall model architecture of TST is shown in Figure 16.

As shown in Figure 16, a new time series tagger is first designed for one-dimensional data processing, and then TST is proposed on this basis in conjunction with the Transformer model. Finally, the CWRU dataset, XJTU-SY dataset, and UCONN dataset were used to verify the validity of the model. Experimental results show that TST achieves 98.63% (ten categories), 99.72% (four categories), 99.78%, and 99.51% fault diagnosis accuracy for the above three datasets, respectively, which are higher than the traditional CNN and RNN models. Meanwhile, the results after feature visualization using *t*-SNE also proved that the feature vectors extracted by TST have the best intra-class closeness and inter-class separability, further proving the effectiveness of the method.

Hou *et al.* [109] proposed a bearing fault diagnosis method based on joint feature extraction of Transformer and ResNet (TAR) for the problems of difficult data acquisition, unbalanced category distribution, and noise interference that often exist in DL models for bearing fault diagnosis driven by big data. The overall model architecture of TAR is shown in Figure 17.

As indicated in Figure 17, feature separation and word embedding were first performed on the original one-dimensional signals through a one-dimensional convolutional layer, which were transmitted to the Transformer encoder and ResNet framework for feature extraction, respectively, and the diagnostic accuracy was better than that of the traditional DL network. In addition, the migration learning strategy employing model fine-tuning reduced the training difficulty of the method in new tasks. Finally, the CWRU dataset was used to verify the validity of the model. The experimental results show that TAR achieves up to 99.90% fault diagnosis accuracy for the CWRU dataset without adding noise, and when adding noise with different signal-to-noise ratios, TAR's average fault diagnosis accuracy is higher than that of the comparison methods.

Fang *et al.* [115] explored a lightweight Transformer based on convolutional embedding and linear self-attention, named CLFormer, for fault diagnosis of rotating machinery. The overall model architecture of CLFormer is shown in Figure 18.

To begin with, the input original one-dimensional signal was normalized in [−1,1], the convolutional embedding module is constructed to replace the original embedding module, to reduce the complexity of the model, the linear self-attention was used to replace the original self-attention, which makes the CLFormer satisfy the demand of lightweight. Finally, the effectiveness of the proposed method was evaluated on a laboratory-measured rotating machinery dataset. The experimental results show that compared with Transformer, the number of parameters of CLFormer decreases from 35.22 to 4.88 K, and the accuracy of fault diagnosis increases from 82.68 to 90.53%, which is of practical application value.

Aiming at the problems of low accuracy and poor robustness of traditional rolling bearing fault diagnosis based on DL, Hou *et al.* [107] designed a multi-feature parallel fusion rolling bearing fault diagnosis method with Transformer as the basic network, named Diagnosisformer. The overall model architecture of Diagnosisformer is shown in Figure 19.

The fast Fourier transform is primarily used to extract the frequency-domain features of the original one-dimensional vibration data, and then the model inputs are subjected to normalization operations and embedded into the network. Then, the multi-feature parallel fusion encoder is used to extract the local and global features of the bearing data, and the corresponding features are passed to the cross-flipped decoder and classified by the classification head for fault classification. Finally, self-made rotating

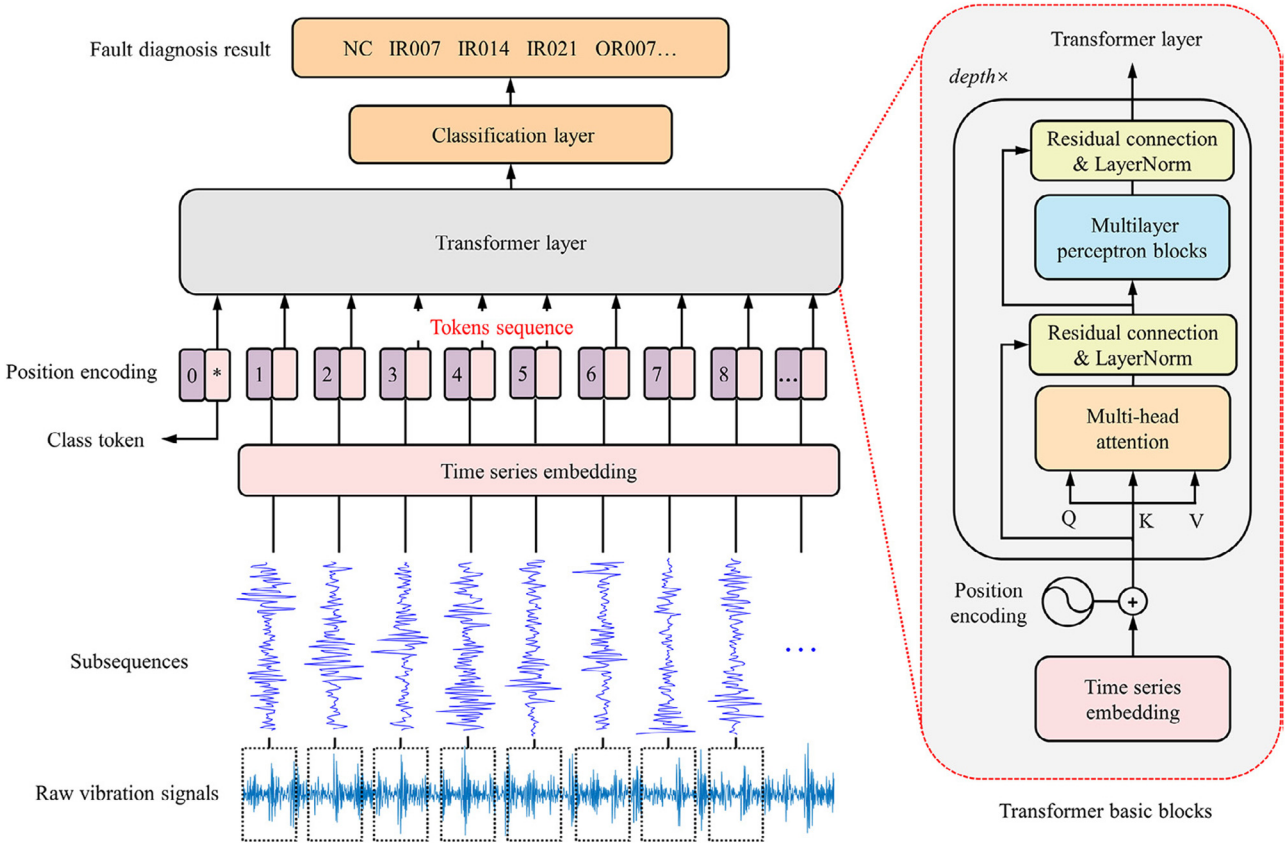

Figure 16: The overall model architecture of TST [106].

machinery fault diagnosis data and the CWRU dataset were used to verify the validity of the model. The experimental results show that Diagnosisformer achieves an average diagnosis accuracy of 99.84 and 99.85% for the above two datasets, respectively. Accuracy and robustness are significantly better than CNN, CNN-LSTM, RNN, LSTM, GRU, and other methods.

Yang *et al.* [108] proposed a signaling Transformer (SiT) on the basis of the attention mechanism and applied it to a study of bearing fault diagnosis. The overall model architecture of SiT is shown in Figure 20.

As demonstrated in Figure 20, the original one-dimensional vibration time series was first segmented, then the segmented subsequence was linearly encoded and positionally encoded, and finally the encoded subsequence was input into the Transformer for feature extraction to realize the fault pattern recognition of bearings. The effectiveness of the method was finally verified using the CWRU bearing dataset and the self-made centrifugal pump bearing dataset. The experimental results show that SiT achieves an average diagnostic accuracy of 99.46 and 99.53% for the above two datasets, respectively, which verifies the effectiveness of the proposed method.

Aiming at the problem that the self-attention mechanism in the current Transformer model can only focus on the correlation information within the sequence and cannot understand the information gap between the samples, Li *et al.* [116] proposed a Twins Transformer model based on two-branch Twins attention for bearing fault diagnosis. The overall model architecture of Twins Transformer is shown in Figure 21.

The proposed Twins Transformer uses cross-attention for the first time to compute correlation information between samples. In addition to retaining the correlation information within the sequence data obtained by computing the self-attention, the cross attention is also utilized to learn the correlation information between the samples. Finally, the performance of the model is validated on four commonly used bearing datasets, the CWRU dataset, UPB dataset [117], MFPT dataset [118], and JUN dataset [119]. The average accuracy of each dataset is improved by 1.73–99.42% compared to the original Transformer.

In terms of visual Transformer applications, Ding *et al.* [120] processed the original one-dimensional vibration signal of rolling bearings *via* synchronous compression WT to obtain a multi-channel time-frequency representation, which was then input into a new time-frequency Transformer (TFT)

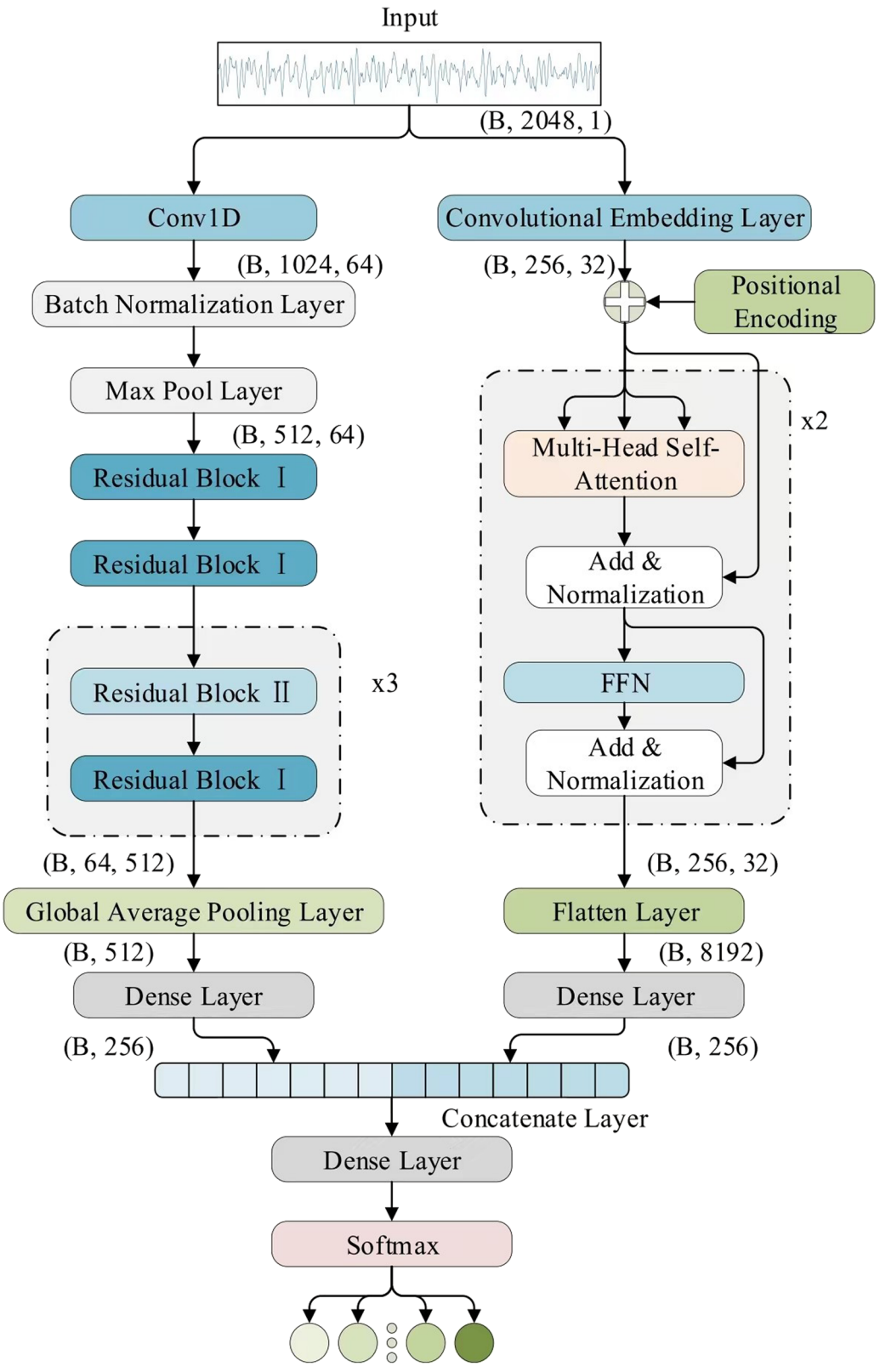

Figure 17: The overall model architecture of TAR [109].

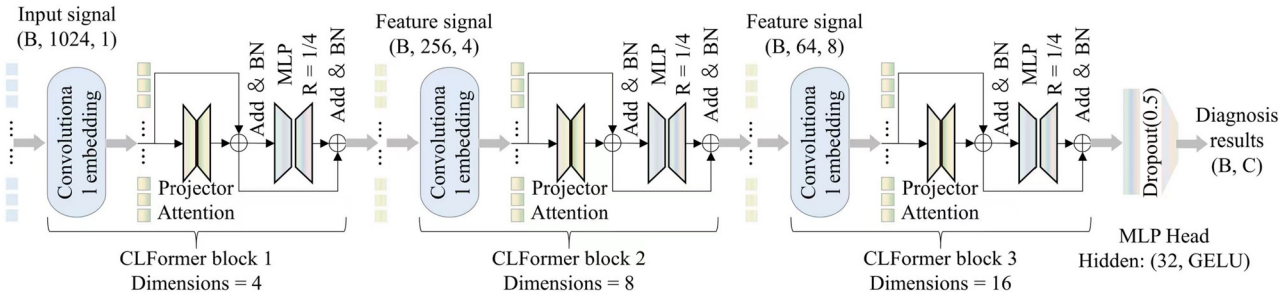

Figure 18: The overall model architecture of CLFormer [115].

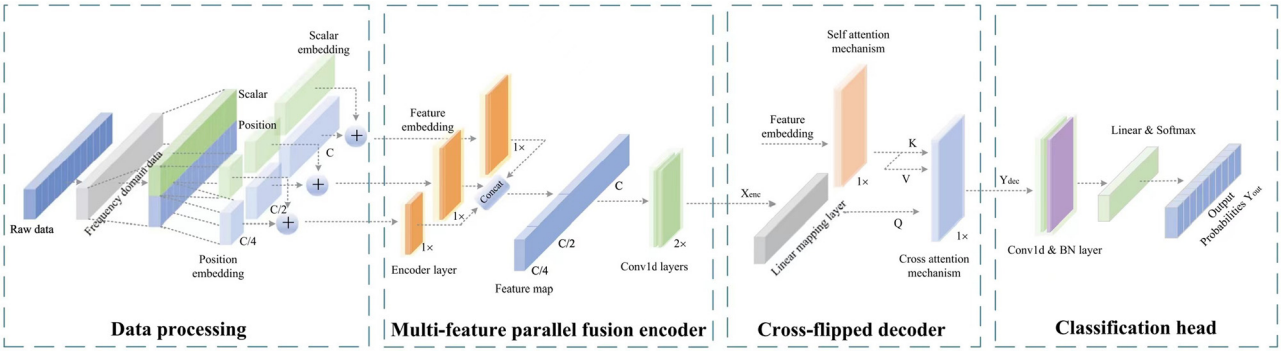

Figure 19: The overall model architecture of Diagnosisformer [107].

to extract discriminative hidden features and accurately classify fault types. The overall model architecture of TFT is shown in Figure 22.

Finally, a self-made bearing experiment dataset was used to verify the validity of the model. Compared with other DL models, this method had higher diagnostic accuracy and faster training speed. Thus, the superiority of the method is demonstrated.

Tang *et al.* [121] explored an integrated ViT model based on WT and soft voting for bearing fault diagnosis. The overall model architecture of the integrated ViT is shown in Figure 23.

First, the original one-dimensional vibration signal was decomposed into sub signals of different frequency bands *via* discrete wavelet transform, and then these sub signals were transformed into wavelet time-frequency maps using CWT. Second, the wavelet time-frequency maps were input into ViT for preliminary diagnostic analysis. Finally, the soft voting method was employed to fuse all preliminary diagnostic results and obtain the final diagnostic decision result. The CWRU dataset was used to verify the validity of the model. The experimental results show that the integrated ViT has the highest fault diagnosis accuracy on all the three datasets, which are 100, 99.67, and 99.83%, respectively, which is better than the integrated CNN and ViT and other comparison methods.

Fan *et al.* [122] proposed a ViT-based fault diagnosis method of rolling bearings to improve the accuracy of rolling bearing fault diagnosis. The overall ViT model architecture is shown in Figure 24.

As illustrated in Figure 24, the original one-dimensional vibration signal was transformed into a grayscale texture image through local binarization, segmented into predetermined sized small blocks, and then transformed into a sequence through linear mapping. The global information of the image was extracted through self-attention mechanism to achieve bearing fault diagnosis. To improve the image recognition performance of ViT, pooling layers were introduced, and the accuracy of the new pooling ViT was improved by 3.3% compared to the original ViT.

Cui *et al.* [123] conducted research on fault diagnosis of waterborne diesel engines under various internal and external excitations, and a fault diagnosis method based on the complementary ensemble EMD of adaptive noise, signal-to-image conversion, and Swin Transformer was proposed. The overall Swin Transformer model architecture is shown in Figure 25.

As is shown in Figure 25, the original one-dimensional vibration signals were first decomposed to obtain a time-frequency matrix. Second, the time-frequency matrix was converted into a two-dimensional color image through

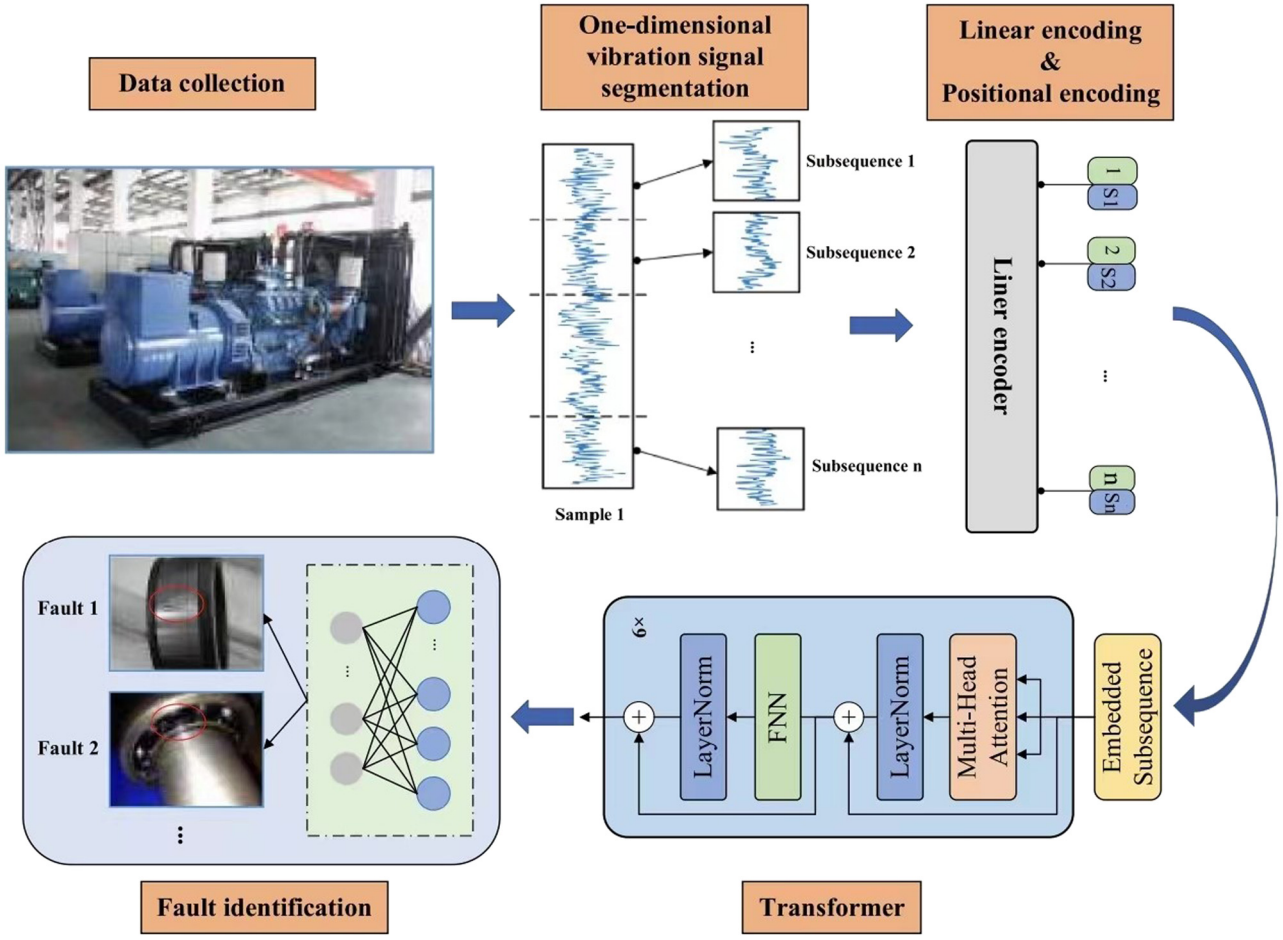

Figure 20: The overall model architecture of SiT [108].

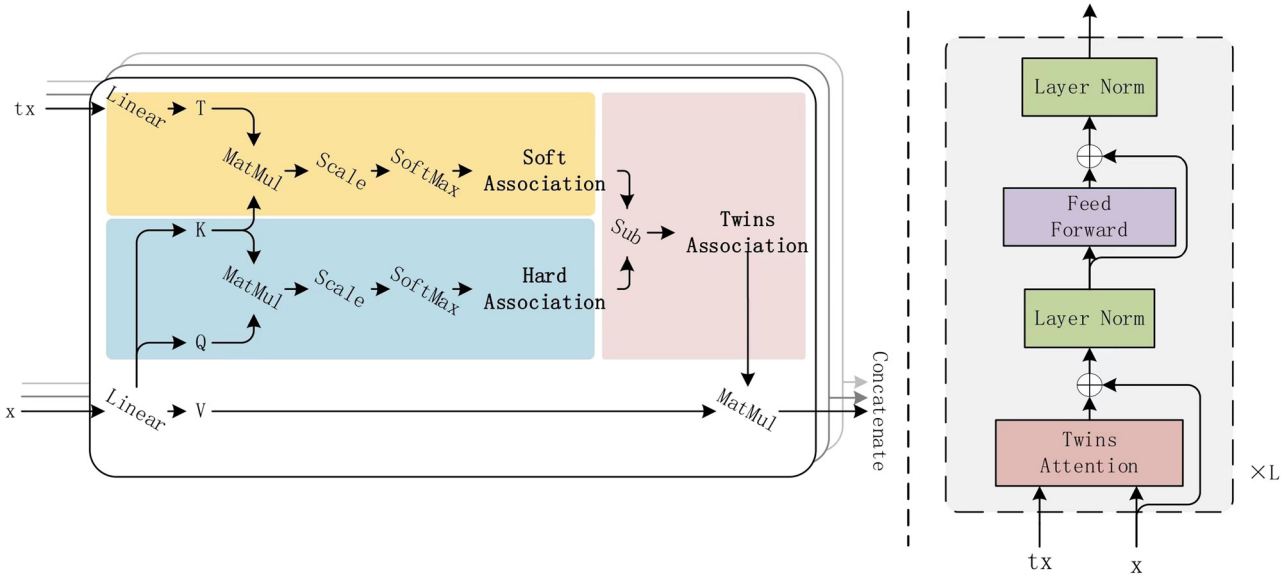

Figure 21: The overall model architecture of Twins Transformer [116].

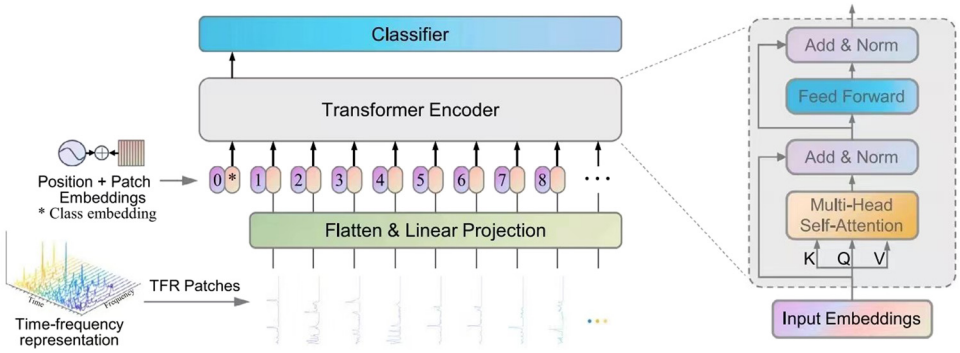

Figure 22: The overall model architecture of TFT [120].

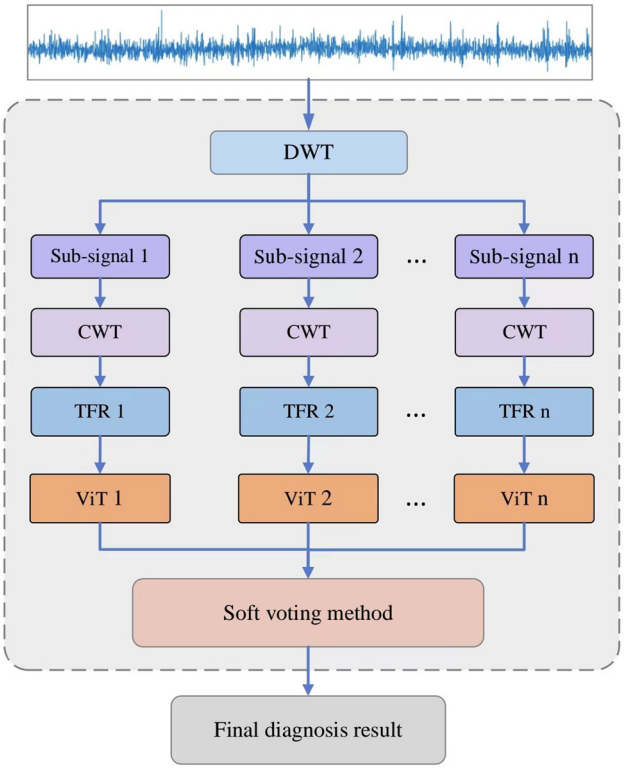

Figure 23: The overall model architecture of the integrated ViT [121].

signal-to-image conversion and pseudocolor encoding. Finally, the two-dimensional image was input into Swin Transformer to fully extract feature information for fault diagnosis. The self-made marine diesel engine dataset was used to verify the effectiveness of the method, and the experimental results proved that the method can effectively extract the fault feature information, and the average fault diagnosis accuracy can reach 98.3%, and has a certain degree of anti-noise interference ability.

Jin *et al.* [124], aiming at the reality that traditional bearing fault diagnosis methods rely on a large amount of manually labeled data, proposed a bearing intelligent fault diagnosis method based on WT and self-supervised learning, which effectively solves the problem of insufficient fault training samples. The overall model architecture is shown in Figure 26.

First, WT and cubic spline difference methods are used to convert the original one-dimensional vibration data into two-dimensional wavelet time-frequency maps, which are then input into the ViT network for feature extraction, KNN is used for fault classification, while the label-free self-distillation algorithm is utilized to solve the problem of self-supervised learning for both finite-labeled data and sufficiently unlabeled data. Finally, the CWRU dataset and XJTU dataset are employed to validate the model. The experimental results show that the method can obtain more than 90% average fault diagnosis accuracy in both datasets with only 1% labeled data, and the comparison results with other self-supervised learning methods also prove the effectiveness and superiority of the method.

He *et al.* [110] proposed a bearing fault diagnosis method Siamese vision Transformer by fusing Siamese network and ViT under the conditions of limited labeled training data and complex working conditions, and the overall model architecture is shown in Figure 27.

Firstly, the STFT is used to convert the one-dimensional vibration signal of the bearing into a two-dimensional time-frequency map, and the feature vectors of the input samples are efficiently extracted in the high-order space to accomplish the fault diagnosis. In the training process of the model, the loss function combining the KL dispersion in both directions and a new random masking strategy are proposed and designed. Finally, the CWRU dataset and Paderborn dataset are used to validate the model. The experimental results show that the proposed method achieves 97.56 and 98.11% average fault diagnosis accuracies, respectively, with limited data, demonstrating the generalization and effectiveness of the method.

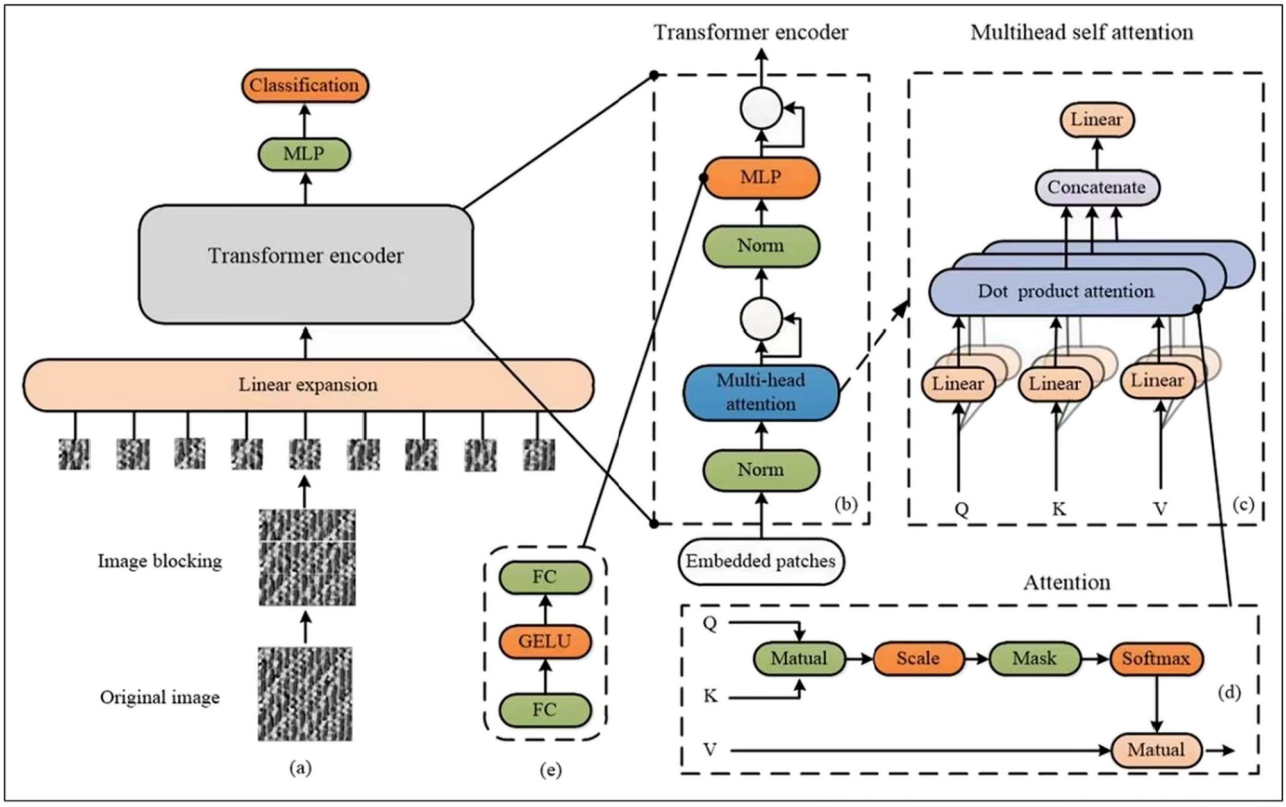

Figure 24: The overall ViT model architecture [122].

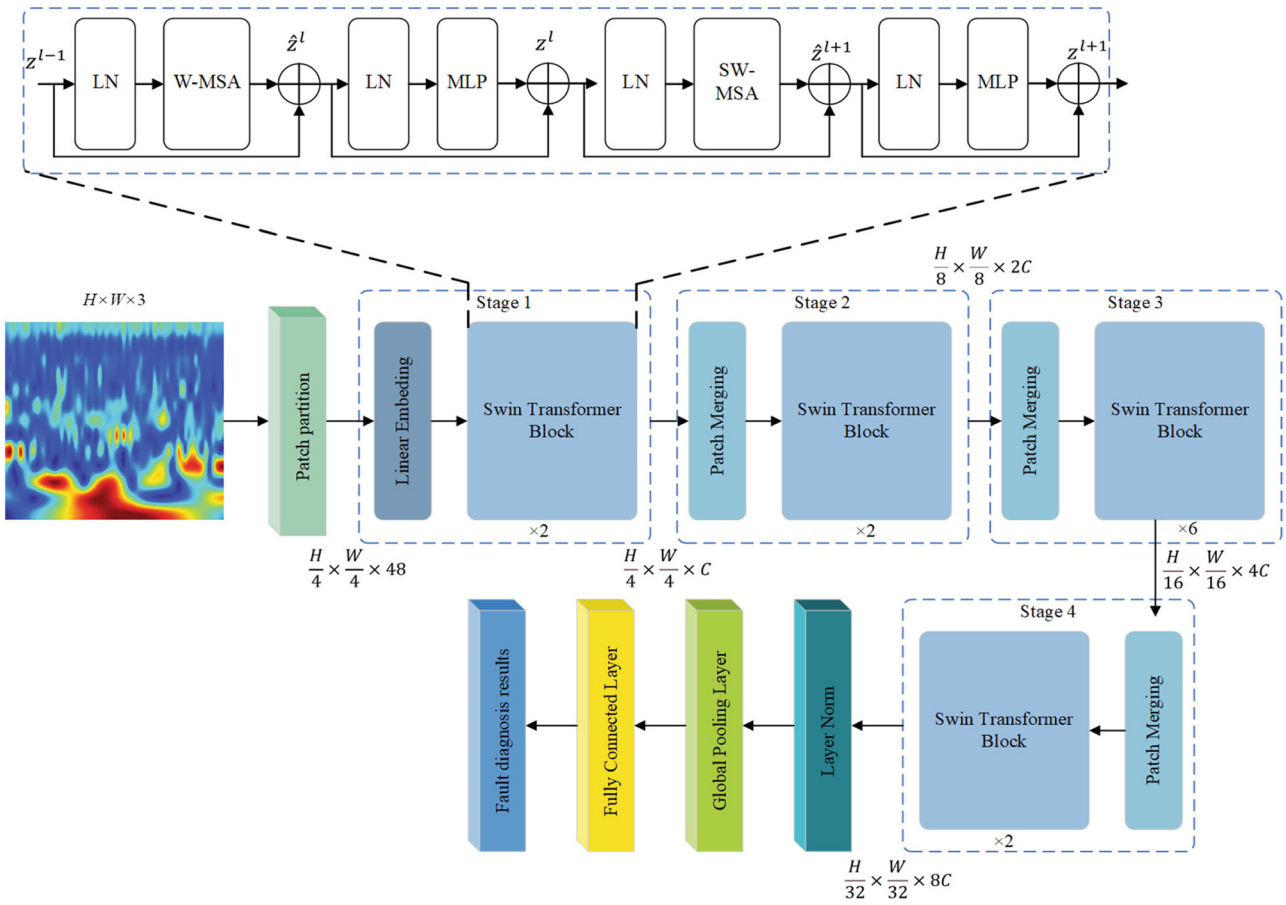

Figure 25: The overall Swin Transformer model architecture [123].

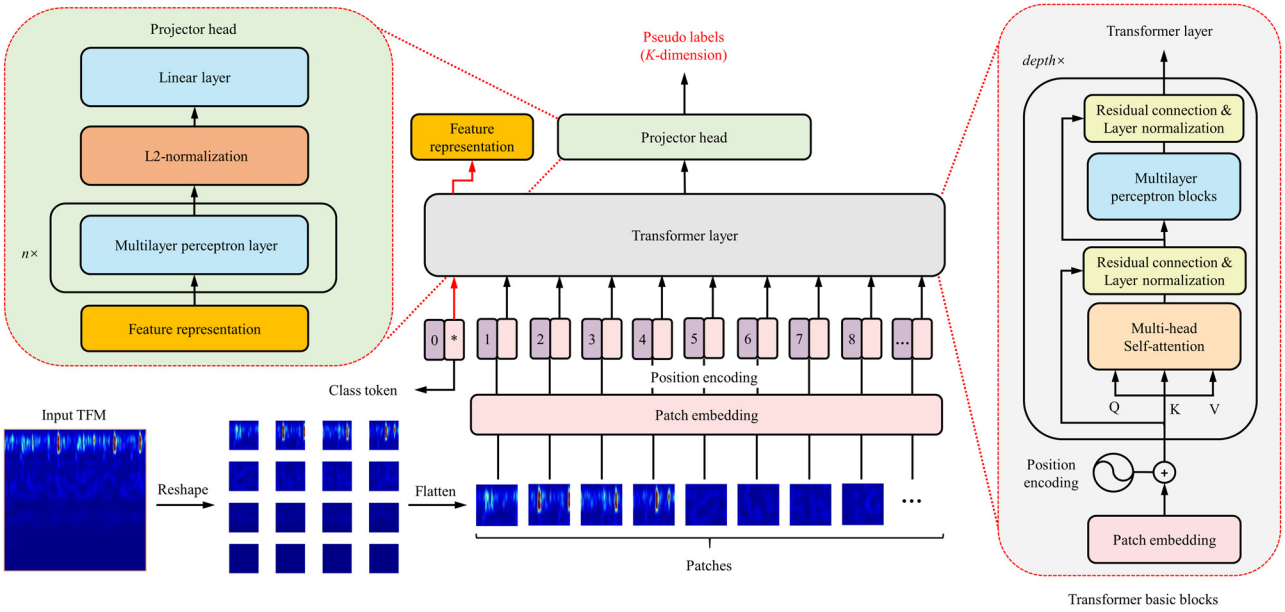

Figure 26: The overall model architecture [124].

3.3.3 Summary of research status

The current research status demonstrates that although some achievements have been made in the research of Transformer-based intelligent fault diagnosis of mechanical equipment, Transformer has attracted the attention of some scholars due to its advantages such as self-attention mechanism, parallel computing, multi-task learning, and flexible and scalable structure. Transformer has gained rapid and wide dissemination in the 6 years since it was proposed, and has made

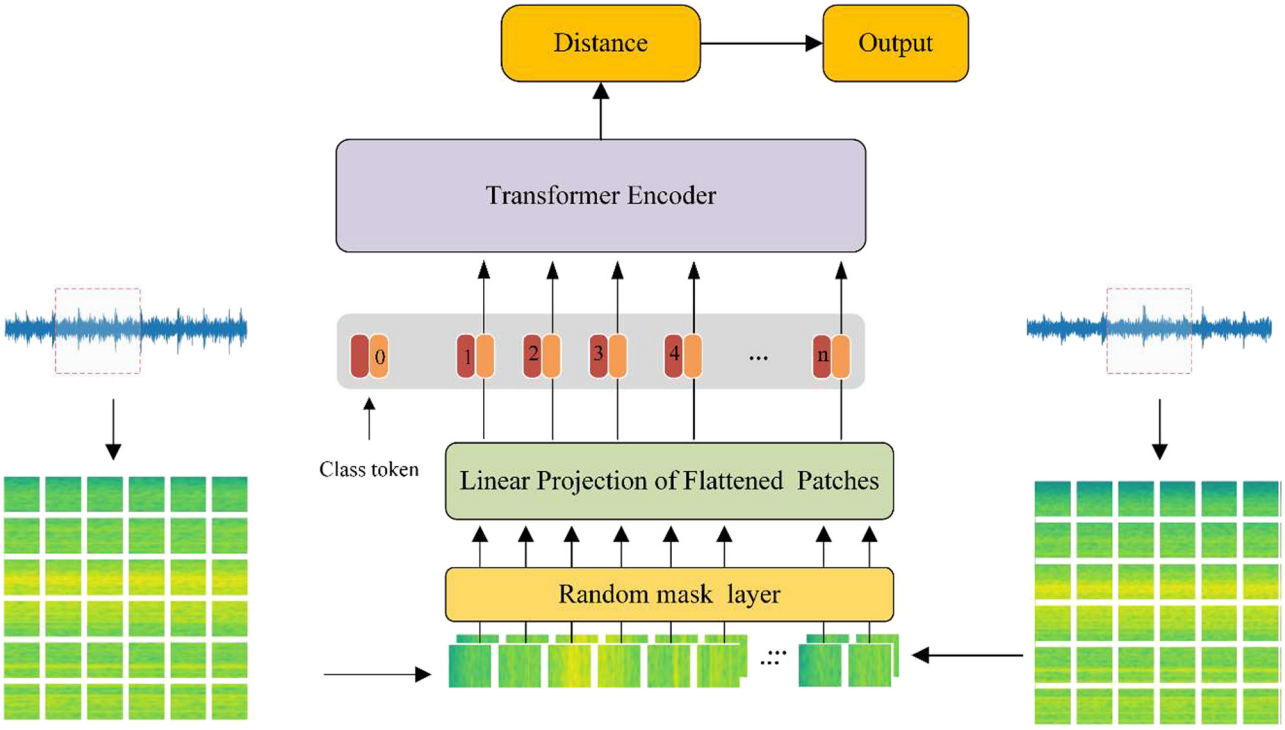

Figure 27: The overall model architecture of Siamese vision Transformer [110].

impressive achievements in many tasks such as NLP, speech recognition, image processing, recommender systems, *etc*. However, so far, the research and applications related to mechanical equipment fault diagnosis using Transformer and its variants are still very limited and have the following problems:

1) The structure of Transformer is relatively complex and contains a large number of parameters, which requires large-scale training samples when pre-training the model and consumes a large number of computational resources and time, increasing the risk of overfitting. For some small datasets or domain-specific fault diagnosis tasks, Transformer may not be able to fully learn the fault features, which leads to poor fault diagnosis results.
2) Currently, most of the Transformer-based mechanical equipment fault diagnosis studies collect mechanical equipment state information under stable operating conditions with fixed operating parameters, environmental variables, *etc*., and many of the studies only analyze single fault patterns. However, the state of mechanical equipment in the actual operation process is mostly a complex fault pattern that occurs across working conditions such as compound speed or multiple faults. Therefore, only the study of stable working conditions or a single fault pattern for fault diagnosis of mechanical equipment does not meet the actual situation.
3) The application of Transformer in the field of mechanical equipment intelligent fault diagnosis is common in rolling bearings, rotor systems, and other common rotating machinery, and there is a great lack of research on typical reciprocating machinery such as diesel engines, and the method validation is mostly based on the public dataset, and there are fewer studies based on the data collected from actual industrial equipment.
4) Research on intelligent fault diagnosis of mechanical equipment based on Transformer has achieved a series of remarkable results. These results not only validate the effectiveness and accuracy of Transformer methods, but also reveal the close relationship between relevant findings and physical phenomena. For example, some studies have found that certain vibration signals of specific frequencies are highly correlated with certain types of mechanical equipment faults. This discovery not only provides a new characteristic index for fault diagnosis, but also provides valuable reference information for the maintenance and management of mechanical equipment. In addition, it has been found that Transformer models can automatically learn some feature representations related to physical phenomena when processing vibration signals. These feature representations not only help to improve the accuracy of fault diagnosis, but also provide a new perspective for understanding the operation mechanism and fault process of mechanical equipment.

4 Research prospects

With the continuous development and improvement of mechanical equipment fault diagnosis methods and techniques, Transformer-based intelligent fault diagnosis of mechanical equipment has become an emerging popular research direction. In recent years, Transformer has been widely applied in various fields of AI, and has become a mainstream method in the fields of CV, NLP, and multimodality. In the future, with the advancement in DL technology, Transformer-based intelligent fault diagnosis methods will be more widely employed. According to the current problems and challenges faced by Transformer-based intelligent fault diagnosis methods of mechanical equipment, this study summarizes and looks forward to its future development direction.

1) **Adaptability to multiple data types.** Currently, Transformer-based intelligent fault diagnosis methods of mechanical equipment are mainly applied to fault diagnosis of rotating machinery such as rolling bearings and rotor systems, and the collected data types are mainly vibration signals and sound signals. In the future, with the development of the industrial Internet, there will be more mechanical equipment operation data, such as temperature, pressure, flow, *etc*., which can be applied for mechanical equipment fault diagnosis, and the scope of application will be further extended to the field of reciprocating machinery such as engines, to further enrich the types and sources of fault diagnosis data, and to improve the reliability of diagnostic results.
2) **Few-shot learning.** Transformer-based diagnostic models usually require massive data samples for training, and then the model is fine-tuned for specific task types to ensure the accuracy and generalization performance of the model. However, it is not realistic to obtain a large amount of labeled data for certain scenarios of mechanical equipment fault diagnosis, because mechanical equipment faults occur randomly, and the sample data are also more complex and difficult to collect and label. Few-shot learning technology can be employed to infer and predict unknown samples from the knowledge learned from a small number of samples by means of algorithm optimization and model improvement, thus solving the problem that Transformer performs well on large datasets but is ineffective in the face of few-shot datasets.

3) **Lightweight research on Transformer.** Transformer model shows excellent performance in mechanical equipment fault diagnosis tasks, due to its high model complexity and large number of parameters, it requires a large amount of computational resource and storage space for training and reasoning, which brings certain challenges to the deployment and application of the model, especially in resource-limited or real-time demanding scenarios, such as mobile devices or embedded systems. Therefore, the lightweight research of Transformer has become an important research direction. Lightweight technology can reduce the complexity of the model by pruning, compressing, and optimizing the model, so as to improve the inference and response speed of the model and reduce the memory occupation of the model, which is of great significance for many real-time and low-latency application scenarios, such as speech recognition, image classification, and real-time fault diagnosis of mechanical equipment.
4) **The fusion development of Transformer and CNN.** Transformer has advantages in processing long sequence data and can effectively capture long-term dependencies. However, it requires a large amount of computing resource and time for training, and its processing performance for short sequence data is not as good as CNN. Meanwhile, CNN performs well in processing local features, especially for data such as images and text, which can effectively capture local features, but it cannot effectively capture global features. Integrating these two models can enable the model to have both global and local feature learning capabilities, thereby improving the representational ability and generalization ability of the model. The fusion development of Transformer and CNN can complement advantages of each other to maximize performance, thus better adapting to different data and tasks.

5 Conclusion

Mechanical equipment fault diagnosis is an important link in modern industrial production, which can effectively prevent catastrophic accidents and reduce economic losses. At present, Transformer and its variants have developed into an effective intelligent fault diagnosis technology for mechanical equipment. This study first reviews the current research status of mechanical equipment fault diagnosis and analyzes the limitations of existing fault diagnosis methods. With the in-depth analysis of four well-known public vibration signal datasets for mechanical equipment faults, the study reveals the intrinsic connection between mechanical equipment fault modes and these physical phenomena, providing a solid physical foundation for intelligent fault diagnosis. On this basis, the network structure and basic principles of Transformer are introduced. Inspired by the successful application of Transformer in the field of NLP, researchers have gradually introduced Transformer into the field of image processing, where image classification is one of the most common applications of the Transformer. Taking this as a starting point, three commonly employed visual Transformer models for image classification are summarized, and a detailed analysis of the research and application of Transformer and their variants in the field of mechanical equipment fault diagnosis are conducted. Finally, the effectiveness and superiority of Transformer and its variants in the field of mechanical equipment intelligent fault diagnosis are further demonstrated through literature comparison, and their future development directions are explored and discussed.

Funding information: The research work was funded by the National Natural Science Foundation of China (No. 71871219).

Author contributions: Conceptualization: Rongcai Wang and Xisheng Jia; methodology: Rongcai Wang and Zichang Liu; writing – original draft preparation: Rongcai Wang; writing – review and editing: Enzhi Dong and Zhonghua Cheng; and supervision: Xisheng Jia. All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Conflict of interest: The authors state no conflict of interest.

Data availability statement: All data generated or analysed during this study are included in this published article.

References

[1] Sun YJ, Wang W. Role of image feature enhancement in intelligent fault diagnosis for mechanical equipment: A review. Eng Fail Anal. 2024;156:107815.
[2] Bently DE, Hatch CT, Grissom B. Fundamentals of rotating machinery diagnostics. New York: ASME Press; 2002.
[3] Jiang DN, Wang ZX. Research on mechanical equipment fault diagnosis method based on deep learning and information fusion. Sensors. 2023;23(15):6999.
[4] Wang CH, Sun YJ, Wang XH. Image deep learning in fault diagnosis of mechanical equipment. J Intell Manuf. 2023.
[5] Yang DL, Zhang WB, Jiang YZ. Mechanical fault diagnosis based on deep transfer learning: A review. Meas Sci Technol. 2023;34(111):112001.

[6] Xie T, Huang X, Choi SK. Intelligent mechanical fault diagnosis using multisensor fusion and convolution neural network. IEEE Trans Ind Inf. 2017;18(5):3213–23.
[7] Du WL, Hu PJ, Wang HC, Gong XY, Wang SY. Fault diagnosis of rotating machinery based on 1D–2D joint convolution neural network. IEEE Trans Ind Electron. 2023;70(5):5277–85.
[8] Zhang W, Zhang T, Cui G, Pan Y. Intelligent machine fault diagnosis using convolutional neural networks and transfer learning. IEEE Access. 2022;10:50959–73.
[9] Liu HY, Ma RZ, Li DY, Yan L, Ma ZM. Machinery fault diagnosis based on deep learning for time series analysis and knowledge graphs. J Sign Process Syst. 2021;93:1433–55.
[10] Zhang Y, Zhou T, Huang X. Fault diagnosis of rotating machinery based on recurrent neural networks. Measurement. 2021;171:108774.
[11] Vaswani A, Shazeer N, Parmar N, Llion J, Gomez AN, Kaiser L, et al. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017. p. 6000–10.
[12] Xu MZ, Yang BS, Wong DF, Chao LS. Multi-view self-attention networks. Knowl-Based Syst. 2022;241:108268.
[13] Xie EZ, Wang WH, Yu ZD, Anandkumar A, Alvarez JM, Luo P SegFormer-Simple and efficient design for semantic segmentation with Transformers. 2021;arXiv:2105.15203.
[14] Chen LC, Kokkinos PI, Murphy K. DeepLab-Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal. 2018;40(4):834–48.
[15] Wu YM, Cao RH, Hu YK, Wang J, Li KL. Combining global receptive field and spatial spectral information for single-image hyperspectral super-resolution. Neurocomputing. 2023;542:126277.
[16] Frank PM, Ding SX, Marcu T. Model-based fault diagnosis in technical processes. Trans Inst Meas Control. 2000;22(1):57–101.
[17] Gao Z, Cecati C, Ding SX. A survey of fault diagnosis and fault-tolerant techniques-part I-fault diagnosis with model-based and signal-based approaches. IEEE Trans Ind Electron. 2015;62(6):3757–67.
[18] Wang XM, Wang J, Privault M. Artificial intelligent fault diagnosis system of complex electronic equipment. J Intell Fuzzy Syst. 2018;35:4141–51.
[19] Pei XL, Zheng XY, Wu JL. Rotating machinery fault diagnosis through a Transformer convolution network subjected to transfer learning. IEEE Trans Instrum Meas. 2021;70:1–11.
[20] Xue F, Zhang WM, Xue F, Li DD, Xie SL, Fleischer J. A novel intelligent fault diagnosis method of rolling bearing based on two-stream feature fusion convolutional neural network. Measurement. 2021;176:109226.
[21] Zhang GY, Wang Y, Li XM, Tang BP, Qin Y. Enhanced symplectic geometry mode decomposition and its application to rotating machinery fault diagnosis under variable speed conditions. Mech Syst Signal Process. 2022;170:108841.
[22] Toutountzakis T, Tan CK, Mba D. Application of acoustic emission to seeded gear fault detection. NDT E Int. 2005;38(1):27–36.
[23] Cheng CC, Yang SY, Lee DS. Novel real-time temperature diagnosis of conventional hot-embossing process using an ultrasonic transducer. Sensors. 2014;14(10):19493–506.
[24] Raposo H, Farinha JT, Fonseca I, Ferreira LA. Condition monitoring with prediction based on diesel engine oil analysis: a case study for urban buses. Actuators. 2019;8(1):14.
[25] Glowacz A, Glowacz Z. Diagnostics of stator faults of the single-phase induction motor using thermal images, MoASoS and selected classifiers. Appl Acoust. 2017;117:20–7.
[26] Choudhary A, Goyal D, Letha SS. Infrared thermography-based fault diagnosis of induction motor bearings using machine learning. IEEE Sens J. 2021;21(2):1727–34.
[27] Duan LX, Yao MC, Wang JJ. Segmented infrared image analysis for rotating machinery fault diagnosis. Infrared Phys Techn. 2016;77:267–76.
[28] Ciabattoni L, Ferracuti F, Freddi A. Statistical spectral analysis for fault diagnosis of rotating machines. IEEE Trans Ind Electron. 2017;65(5):4301–10.
[29] Yang MS, Sinaga KP. Collaborative feature-weighted multi-view fuzzy c-means clustering. Pattern Recogn. 2021;119:108064.
[30] Seo DH, Choi JW, Kim YH. Impulsive sound source localization using peak and RMS estimation of the time-domain beamformer output. Mech Syst Signal Pr. 2014;49(1–2):95–105.
[31] Liao L. Discovering prognostic features using genetic programming in remaining useful life prediction. IEEE Trans Ind Electron. 2014;61(5):2464–72.
[32] Figlus T. A method for diagnosing gearboxes of means of transport using multi-stage filtering and entropy. Entropy. 2019;21(5):441.
[33] Li Y, Cheng G, Ma SC, Li X. Bearing fault diagnosis method based on complete center frequency distribution feature. Struct Health Monit. 2023;22(6):4100–16.
[34] Ho CYF, Ling BWK, Deng DX, Liu YW. Tachycardias classification via the generalized mean frequency and generalized frequency variance of electrocardiograms. Circuits Syst Signal Process. 2022;41:1207–22.
[35] Kim JG, Mukherjee S, Bates A, Zickel B, Park S, Son BR, et al. Autocorrelation standard deviation and root mean square frequency analysis of polymer electrolyte membrane fuel cell to monitor for hydrogen and air undersupply. J Power Sources. 2015;300:164–74.
[36] Zhang SQ, Sun YF, Dong W, You SZ, Liu YZ. Diagnosis of bearing fault signals based on empirical standard autoregressive power spectrum signal decomposition method. Meas Sci Technol. 2023;35(1):015010.
[37] Jiang F, Ding K, He GL, Du CY. Sparse dictionary design based on edited cepstrum and its application in rolling bearing fault diagnosis. J Sound Vib. 2021;490:115704.
[38] Wang LH, Zhao XP, Wu JX, Xie YY, Zhang YH. Motor fault diagnosis based on short-time Fourier transform and convolutional neural network. Chin J Mech Eng. 2017;30(6):1357–68.
[39] Cheng C, Zhao ZY, Tang C, Qian GC, Islam S. Diagnostic of Transformer winding deformation fault types using continuous wavelet transform of pulse response. Measurement. 2019;140:197–206.
[40] Ma ZQ, Ruan WY, Chen MY, Li X. An improved time-frequency analysis method for instantaneous frequency estimation of rolling bearing. Shock Vib. 2018;2018:8710190.
[41] Jin S, Lee SK. Bearing fault detection utilizing group delay and the Hilbert-Huang transform. J Mech Sci Technol. 2017;31(3):1089–96.
[42] Ma ZY, Lai YP, Xie JY, Meng DY, Kleijn WB, Guo J, et al. Dirichlet process mixture of generalized inverted Dirichlet distributions for positive vector data with extended variational inference. IEEE Trans Neural Netw Learn Syst. 2022;33(11):6089–102.
[43] Zhang XX, Chen DG, Wu KS. Incremental nonnegative matrix factorization based on correlation and graph regularization for matrix completion. Int J Mach Learn Cyber. 2019;10:1259–68.
[44] Yan DY, Li KP, Gu S, Yang L. Network-based Bag-of-Words model for text classification. IEEE Access. 2020;8:82641–52.

[45] Chen KW, Zhang ZP, Long J, Zhang H. Turning from TF-IDF to TF-IGM for term weighting in text classification. Expert Syst Appl. 2016;66:245–60.
[46] Sharma A, Kumar S. Ontology-based semantic retrieval of documents using Word2vec model. Data Knowl Eng. 2023;144:102110.
[47] Mahto D, Yadav SC. Emotion prediction for textual data using GloVe based HeBi-CuDNNLSTM model. Multimed Tools Appl. 2024;83:18943–68.
[48] Gao L, Li DH, Yao LL, Gao YN. Sensor drift fault diagnosis for chiller system using deep recurrent canonical correlation analysis and k-nearest neighbor classifier. ISA Trans. 2022;122:232–46.
[49] Chen MQ, Qu R, Fang WG. Case-based reasoning system for fault diagnosis of aero-engines. Expert Syst Appl. 2022;202:117350.
[50] Sun WX, Chen J, Li JQ. Decision tree and PCA-based fault diagnosis of rotating machinery. Mech Syst Signal Process. 2007;21:1300–17.
[51] Amarnath M, Sugumaran V, Knmar H. Exploiting sound signals for fault diagnosis of bearings using decision tree. Measurement. 2013;46:1250–6.
[52] Sakthivel NR, Sugumaran V, Babudevasenapati S. Vibration based fault diagnosis of monoblock centrifugal pump using decision tree. Expert Syst Appl. 2010;37:4040–9.
[53] Cerrada M, Zurita G, Cabrera D, Sánchez RV, Artés M, Li C. Fault diagnosis in spur gears based on genetic algorithm and random forest. Mech Syst Signal Process. 2016;70–71:87–103.
[54] Li C, Sanchez RV, Zurita G, Cerrada M, Cabrera D, Vásquez RE. Gearbox fault diagnosis based on deep random forest fusion of acoustic and vibratory signals. Mech Syst Signal Process. 2016;76–77:283–93.
[55] Hu Q, Si XS, Zhang QH, Cerrada M, Cabrera D, Vásquez RE. A rotating machinery fault diagnosis method based on multi-scale dimensionless indicators and random forests. Mech Syst Signal Process. 2020;139:106609.
[56] Cheng G, Li HY, Hu X, Chen XH, Liu HG. Fault diagnosis of gearbox based on local mean decomposition and discrete hidden Markov models. Proc Inst Mech Eng C-J Mech Eng Sci. 2017;231(14):2706–17.
[57] Lopez G, Naranjo A, Lu SL, Moore KJ. Hidden Markov model based stochastic resonance and its application to bearing fault diagnosis. J Sound Vib. 2022;528:116890.
[58] Deng F, Guo S, Zhou R, Chen J. Sensor multifault diagnosis with improved support vector machines. IEEE Trans Autom Sci Eng. 2017;14(2):1053–63.
[59] Ma YS, Yao JN, Ma C, Xiao XM. Pattern recognition of rigid hoist guides based on support vector machine. Adv Mech Eng. 2018;10(12):1–7.
[60] Wang Y, Huang YH, Yang K, Chen ZH, Luo C. Generator fault classification method based on multi-source information fusion naive Bayes classification algorithm. Energies. 2022;15(24):9635.
[61] An ZN, Wu F, Zhang C, Ma JH, Sun B, Tang BH, et al. Deep learning-based composite fault diagnosis. IEEE J Em Sel Top C. 2023;13(2):572–81.
[62] Hoang DT, Kang HJ. A survey on Deep Learning based bearing fault diagnosis. Neurocomputing. 2019;335:327–35.
[63] He M, He D. Deep learning based approach for bearing fault diagnosis. IEEE Trans Ind Appl. 2017;53(3):3057–65.
[64] Ma S, Chu FL. Ensemble deep learning-based fault diagnosis of rotor bearing systems. Comput Ind. 2019;105:143–52.
[65] Chen YJ, Rao M, Feng K, Niu G. Modified varying index coefficient autoregression model for representation of the nonstationary vibration from a planetary gearbox. IEEE Trans Instrum Meas. 2023;72:1–12.
[66] Chen YJ, Rao M, Feng K, Zuo MJ. Physics-Informed LSTM hyperparameters selection for gearbox fault detection. Mech Syst Signal Process. 2022;171:108907.
[67] Han T, Xie WZ, Pei ZY. Semi-supervised adversarial discriminative learning approach for intelligent fault diagnosis of wind turbine. Inf Sci. 2023;648:119496.
[68] Yao JH, Han T. Data-driven lithium-ion batteries capacity estimation based on deep transfer learning using partial segment of charging/discharging data. Energy. 2023;271:127033.
[69] Go Z, Yang J, Zhang L, Jiang FL, Jiao XX. TEGAN: Transformer embedded generative adversarial network for underwater image enhancement. Cogn Comput. 2024;16:191–214.
[70] Naveen S, Kiran MSSR, Indupriya M, Manikanta TV, Sudeep PV. Transformer models for enhancing AttnGAN based text to image generation. Image Vis Comput. 2021;115:104284.
[71] Hu X, Li T, Zhou T, Liu Y, Peng YX. Contrastive learning based on Transformer for hyperspectral image classification. Appl Sci-Basel. 2021;11(18):8670.
[72] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S End-to-end object detection with Transformers. 2020;arXiv:2005.12872.
[73] He KK, Gou FF, Wu J. Image segmentation technology based on Transformer in medical decision-making system. IET Image Process. 2023;17(10):3040–54.
[74] Parmar N, Vaswani A, Uszkoreit J, Kaiser Ł, Shazeer N, Ku A, et al. Image Transformer. 2018;arXiv:1802.05751.
[75] Dosovitskiy A, Beyer L, KolesnikovI A, Weissenborn D, Zhai XH, Unterthiner T, et al. An image is worth 16 × 16 words-Transformers for image recognition at scale. 2020;arXiv:2010.11929.
[76] Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H Training Data-efficient Image Transformers & distillation through attention. 2020;arXiv:2012.12877.
[77] Liu Z, Lim YT, Cao Y, Hu H, Wei YX, Zhang Z, et al. Transformer: hierarchical vision Transformer using shifted windows. 2021;arXiv:2103.14030.
[78] Bi JR, Zhu ZL, Meng QL. Transformer in computer vision. 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology; 2021. p. 178–88.
[79] Jonas G, Michael A, David G, Yarats D, Dauphin YN Convolutional sequence to sequence learning. 2017;arXiv:1705.03122.
[80] Niu ZY, Zhong GQ, Hui Y. A review on the attention mechanism of deep learning. Neurocomputing. 2021;452:48–62.
[81] Lieskovska E, Jakubec M, Jarina R, Chmulík M. A review on speech emotion recognition using deep learning and attention mechanism. Electronics. 2021;10(10):1163.
[82] Choi H, Cho KY, Bengio YS. Fine-grained attention mechanism for neural machine translation. Neurocomputing. 2018;284:171–6.
[83] Wang W, Shen J, Yu Y, Ma KL. Stereoscopic thumbnail creation via efficient stereo saliency detection. IEEE Trans Vis Comput Gr. 2016;23(8):2014–27.
[84] Lin Z, Feng M, Santos CND, Yu M, Xiang B, Zhou B, et al. A structured self-attentive sentence embedding. 2017;arXiv:1703.03130.
[85] Han K, Xiao A, Wu EH, Guo JY, Xu CJ, Wang YH. Transformer in Transformer. 2021;arXiv:2103.00112.

[86] Wang WH, Xie EZ, Li X, Fan DP, Song KT, Liang D, et al. Pyramid vision Transformer: a versatile backbone for dense prediction without convolutions. 2021;arXiv:2102.12122.
[87] Yuan L, Chen YP, Wang T, Yu WH, Shi YJ, Jiang ZH, et al. Tokens-to-token ViT: training vision Transformers from scratch on Imagenet. 2021;arXiv:2101.11986.
[88] Zhou DQ, Kang BY, Jin XJ, Yang LJ, Lian XC, Jiang ZH, et al. DeepViT: towards deeper vision Transformer. 2021;arXiv:2103.11886.
[89] Touvrom H, Cord M, Sablayrolles A, Synnaeve G, Jégou H. Going deeper with Image Transformers. 2021;arXiv:2103.17239.
[90] Chen CFR, Fan QF, Panda R. CrossViT: cross-attention multi-scale vision Transformer for image classification. 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021. p. 347–56.
[91] Sun ZQ, Cao SC, Yang YM, Kitani Q. Rethinking Transformer-based set prediction for object detection. 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021. p. 3591–600.
[92] Zheng MH, Gao P, Zhang RR, Li KC, Wang XG, Li HS, et al. End-to-end object detection with adaptive clustering Transformer. 2020;arXiv:2011.09315.
[93] Zhang D, Zhang HW, Tang JH Feature Pyramid Transformer. 2020;arXiv:2007.09451.
[94] Zheng SX, Lu JC, Zhao HS, Zhu XT, Luo ZK, Wang YB, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with Transformers. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021. p. 6877–86.
[95] Strudel R, Garcia R, Laptev I, Schmid C. Segmenter: Transformer for semantic segmentation. 2021;arXiv:2105.05633.
[96] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 770–8.
[97] Tan MX, Le QV. Efficientnet: rethinking model scaling for convolutional neural networks. 2019;arXiv:1905.11946.
[98] Han K, Wang YH, Chen HT, Chen XH, Guo JY, Liu ZH, et al. A survey on Vision Transformer. IEEE Trans Pattern Anal. 2023;45(1):87–110.
[99] Case Western Reserve University Bearing Data Center; 2021; https://engineering.case.edu/bearingdatacenter.
[100] Wang B, Lei YG, Li NP, Li NB. A hybrid prognostics approach for estimating remaining useful life of rolling element bearings. IEEE Trans Reliab. 2020;69(1):401–12.
[101] Cao P, Zhang S, Tang J. Preprocessing-free gear fault diagnosis using small datasets with deep convolutional neural network-based transfer learning. IEEE Access. 2018;6:26241–53.
[102] Hou L, Yi HM, Jin YH, Gui M. Inter-shaft bearing fault diagnosis based on aero-engine system: a benchmarking dataset study. J Dynam Monit Diagn. 2023;2(4):228–42.
[103] Jia F, Lei YG, Lin J, Zhou X, Lu N. Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. Mech Syst Signal Process. 2016;72-73:303–15.
[104] Shao HD, Jiang HK, Zhang X, Niu MG. Rolling bearing fault diagnosis using an optimization deep belief network. Meas Sci Technol. 2015;26(11):115002.
[105] Eren L, Ince T, Kiranyaz S. A generic intelligent bearing fault diagnosis system using compact adaptive 1D CNN classifier. J Sig Process Syst. 2019;91:179–89.
[106] Jin HY, Hou L, Chen YS. A Time Series Transformer based method for the rotating machinery fault diagnosis. Neurocomputing. 2022;494:379–95.
[107] Hou YD, Wang JJ, Chen ZQ, Ma JL, Li TJ. Diagnosisformer: An efficient rolling bearing fault diagnosis method based on improved Transformer. Eng Appl Artif Intel. 2023;124:106507.
[108] Yang ZH, Cen J, Liu X, Xiong JB, Chen HH. Research on bearing fault diagnosis method based on Transformer neural network. Meas Sci Technol. 2022;33(8):085111.
[109] Hou SX, Lian H, Chu YD. Bearing fault diagnosis method using the joint feature extraction of Transformer and ResNet. Meas Sci Technol. 2023;34(7):075108.
[110] He QC, Li SB, Bai Q, Zhang AS, Yang J, Shen MM. A Siamese Vision Transformer for bearings fault diagnosis. Micromachines. 2022;13(10):1656.
[111] Wang ZJ, Zheng LK, Wang JY, Du WH. Research on novel bearing fault diagnosis method based on improved krill herd algorithm and kernel extreme learning machine. Complexity. 2019;2019:1–9.
[112] Zhao YJ, Zhou ML, Xu XZ, Zhang NN, Zhang HB. Fault diagnosis based on space mapping and deformable convolution networks. IEEE Access. 2020;8:212599–607.
[113] Zhao ZB, Li TF, Wu JY, Sun C, Wang SB, Yan RQ, et al. Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study. ISA Trans. 2020;107:224–55.
[114] Xu Y, Li ZX, Wang SQ, Li WH, Li WH, Thompson SG, et al. A hybrid deep-learning model for fault diagnosis of rolling bearings. Measurement. 2021;169:108502.
[115] Fang HR, Deng J, Bai YX, Feng B, Li S, SY S, et al. CLFormer: A lightweight Transformer based on convolutional embedding and linear self-attention with strong robustness for bearing fault diagnosis under limited sample conditions. IEEE Trans Instrum Meas. 2022;71:1–8.
[116] Li J, Bao Y, Liu WX, Wang LK, Wang ZB. Twins Transformer: Cross-attention based two-branch Transformer network for rotating bearing fault diagnosis. Measurement. 2023;223:113687.
[117] Lessmeier C, Kimotho JK, Zimmer D, Sextro W. Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. PHM Society European Conference; 2016.
[118] Bechhoefer DE. Fault data sets - society for machinery failure prevention technology; Oct 2012. http://mad-net.org:8765/explore.html?t=0.41789510832145527. Last accessed on 2023-09-10.
[119] Li K. Fault data sets from Jiangnan University; Oct 2012. https://www.mfpt.org/fault-data-sets/, Last accessed on 2023-09-10.
[120] Ding YF, Jia MP, Miao QH, Cao YD. A novel time-frequency Transformer based on self-attention mechanism and its application in fault diagnosis of rolling bearings. Mech Syst Signal Pr. 2022;168:108616.
[121] Tang XY, Xu ZB, Wang ZG. A novel fault diagnosis method of rolling bearing based on integrated Vision Transformer model. Sensors. 2022;22(10):3878.
[122] Fan HW, Ma NG, Zhang XH, Xue CY, Ma JT, Yan Y. New intelligent fault diagnosis approach of rolling bearing based on improved vibration gray texture image and vision Transformer. Proc Inst Mech Eng C-J Mech Eng Sci. 2022.
[123] Cui DX, Hu YH. Fault diagnosis for marine two-stroke diesel engine based on CEEMDAN-Swin Transformer algorithm. J Fail Anal Prev. 2023;23:988–1000.
[124] Jin YH, Hou L, Du M, Chen YS. A Wavelet Transform and self-supervised learning-based framework for bearing fault diagnosis with limited labeled data. 2022;arXiv.2207.10432.