Periodica Polytechnica Electrical Engineering and Computer Science, 68(1), pp. 74–93, 2024

Data Augmented of Mechanical Fault Sound Signal based on Generative Adversarial Networks

Yining Yang[1], Xiang Su[2], Nan Li[2*]

[1] Department of Transportation Engineering, Shanxi Engineering Vocational College, 131 Xinjian Road, 030009 Taiyuan, Shanxi, China
[2] School of Artificial Intelligence, Beijing Technology and Business University, No. 33 Fucheng Road, Haidian District, 100048 Beijing, China
[*] Corresponding author, e-mail: linan@th.btbu.edu.cn

Received: 17 April 2023, Accepted: 28 July 2023, Published online: 20 September 2023

Abstract

In this paper, a global average pooling convolutional neural network based on CNN is proposed for mechanical fault sound detection, which called as GCMD. To solve the data scarcity of mechanical fault sound data, a spectrum frame selection augmented method based on log Mel spectrum feature is proposed to augment the original data, that aim is to train GCMD and generate counter networks. In order to solve the unbalance problem of data set and further improve the generalization ability of GCMD, an augmented neural network model based on CapsuleGAN was proposed, which called MFS-CapsuleGAN. The model was evaluated on the augmented data set by training GCMD neural network. Compared with the original data set, the accurate recognition rate of the model was improved by 23.7%. The performance of this method is improved significantly, which proves the feasibility and effectiveness of MFS-CapsuleGAN data augmented. In addition, the data set with background noise was used to test the generalization ability of GCMD network. The fluctuation range was within 0.117, indicating the good robustness of GCMD network.

Keywords

failure diagnosis, data augment, generating antagonism network, capsule generative adversarial network

1 Introduction

In industry, traditional and intelligent detection are two common detection methods to preliminarily determine whether a machine is malfunctioned. Among them, the auscultation method is the maintenance or testing personnel to analyze the noise frequency band of 20 Hz to 20 kHz audible sound signal. According to the general characteristics of the fault sound of different mechanical parts, the running state of the machine is judged according to the sound of the machine in the running process, the air path is unobstructed, etc., the existence of the fault and the source of the diagnosis of the detection method. However, the traditional artificial experience detection method not only has a serious dependence on people, but also has low efficiency, low accuracy, high cost, which has not met the needs of modern production automation.

At present, the machine intelligent fault detection method generally needs multiple sensors when it detects the fault of the machine. These sensors often need to be in contact with the machine, whether they are connected to the inside or outside of the machine, which may affect the operating conditions of the machine. To improve detection efficiency and reduce the interference of sensor detection, many intelligent detection methods based on neural network have been proposed.

Natural language processing has been widely used in RNN [1], while CNN has less application in sound recognition and classification. However, recent research finds that the development of CNN can be used for reference and help in the sounds field [2]. The effect of CNN on machine learning of sounds type is also quite remarkable.

Many applications of voice recognition are getting closer to people's daily life, which is mainly reflected in modern home intelligent devices. Many applications of voice recognition are getting closer and closer to people's daily life, which is mainly reflected in modern household intelligent devices, such as washing machine noise recognition. Sound detection has been applied in many aspects, including sounds security monitoring [3], urban sound

Cite this article as: Yang, Y., Su, X., Li, N. "Data Augmented of Mechanical Fault Sound Signal based on Generative Adversarial Networks", Periodica Polytechnica Electrical Engineering and Computer Science, 68(1), pp. 74–93, 2024. https://doi.org/10.3311/PPee.22427

analysis [4], machine intelligent fault diagnosis [5] and the bird sound detection challenges which is organized by Queen Mary University of London [6].

However, with the development of technology, the deep learning method has proved to be more effective than the artificial neural network. The intelligent sounds signal fault diagnosis system based on deep learning is a fault diagnosis method based on neural network, which relies heavily on the training data to improve the performance of the model. The small amount of data and the imbalance of positive and negative samples will lead to the phenomenon of network training overfitting or network crash, resulting in the unqualified training model. Usually, the fault sound of different mechanical equipment is only suitable for the training of this kind of sound fault diagnosis deep learning model. The non-general applicability of fault sounds, the number of data species and the diversity of data within the class limit the improvement of neural network classification performance in the application of mechanical sounds fault diagnosis.

There are two problems in data acquisition as follows. First of all, for a class of machinery, the probability of failure is small, and the amount of data that can be collected is far from the amount required by the neural network. Secondly, the efficiency of manual data collection and annotation is very low, which makes it difficult to complete data collection in a short time.

This paper focuses on the fact that the data labels are invariant. In order to solve the data set imbalance and further improve the generalization ability of GCMD, through the comparison of the theoretical performance of GANs, this paper proposes a mechanical fault sound data augmentation neural network model MFS-CapsuleGAN based on CapsuleGAN [7]. CapsuleGAN has the unique learning ability of features and features, as well as the connection between features and the whole, and can analyze the spatial location of the feature data distribution of fault sounds. Compared with DCGAN, CapsuleGAN has the advantage of generating more regular spectrum data, reducing the clutter of synthesized sounds signal characteristics, and improving the complexity of data synthesized data. In addition, small data requirements are also one of the characteristics that distinguish it from DCGAN. Compared with DCGAN, it is easier to satisfy the training conditions. Based on the theory of CapsuleGAN, this model tries to apply CapsuleGAN to the data reinforcement of the original log Mel frequency spectrum of the mechanical fault tone, and synthesize the fault sounds data with the diversity of the original data samples by generating the semi-supervised learning method generated by the confrontation network. MFS-CapsuleGAN can synthesize multi-second sounds log Mel frequency spectrum fragments with global coherence for mechanical fault sounds, and produce log Mel frequency spectrum features similar to the real mechanical fault sound, thus producing appropriate and similar sound effects and better adapting to the generation of mechanical fault sounds. This paper takes the data augmented study of the faulty sound data set of washing machine [8], ToyADMOS [9] and MIMII [10] data set as an example and USES MFS-CapsuleGAN to enhance the original data. The experiment generated the augmented datasets of three kinds of disclosed mechanical fault audio datasets, and the accuracy of the augmented datasets was evaluated by training GCMD neural network. In addition, the data set with background noise was used to test the generalization ability of GCMD network and verify the robustness of GCMD network.

Our contributions are summarized as follows:

1. We propose a method of feature data sampling based on the spectrum feature of log Mel. This method operates on the spectrum of log Mel of a 10 min fault tone. By randomly or orderly selecting the spectrum segment for sampling, we can make the most of the features at the joint of the sample and the sample, and repeat sampling while avoiding feature omission.
2. We proposed MFS-CapsuleGAN. After the improvement of CapsuleGAN network, the log Mel spectrum characteristics of mechanical fault sound signal were enhanced. The end-to-end fault audio synthesis is realized by combining Griffin-Lim algorithm with MFS-CapsuleGAN.

2 Related work

2.1 Data augmentation

To solve the problems about data acquisition, the solution we adopt is data augmentation. That is, the method of expanding the number and diversity of samples by making moderate changes to the in-class sample data while keeping the same label. In the aspect of intelligent fault sound diagnosis and recognition based on deep learning, most of the projects generally have the problem of insufficient training data or unbalanced training set, and the traditional data augmentation methods cannot fully meet the data demand of intelligent fault sound diagnosis and recognition. The generative antagonistic neural network proposed by Ian Goodfellow has been studied by scholars in different fields around the world [11–13]. It is currently an internationally recognized data-augmented deep learning method. This technology is an unsupervised machine learning artificial intelligence algorithm, which can be used to generate unknown new data close to the real data distribution. The framework consists of two neural networks composed of discriminator and generator, which are similar to counterfeiters and truth experts. The generator generates a series of false data, and the discriminator identifies the degree of authenticity of the data and returns the results to the generator. The generator learns in the direction of getting closer to the real data distribution, iteratively updates the parameters and optimizes the model algorithm until the discriminator fails to identify the criterion of authenticity of the data be generated.

In the field of image, many research works of data augmentation technology based on GAN in computer vision have been greatly recognized. GAN is a data augmentation method of supervised learning proposed by Goodfellow et al. [14]. These models take the generative adversarial networks as the framework, obtain the data distribution through discriminator and generator game learning, and generalize it to generate new data in the class. This generation process does not depend on the class itself, but can be applied to unknown new data classes. The generated data not only increases the overall number, but also combines or erases some characteristics of some samples, which makes the generated samples more diverse and more suitable for the learning improvement of model generalization ability.

Based on the network architecture of GAN, Sun et al. [15] proposed a data augmentation method of AC-WGAN-GP to expand the training set and achieve better classification accuracy with a small number of labeled samples. Radford et al. [16] proposed a deep convolution generation countermeasure network DCGAN. By constructing discriminators and generators based on convolution and deconvolution neural networks, the data augmentation performance of GAN is greatly improved. To increase the accuracy of speech recognition, Zhou and Sun [17] suggest a dual data augmentation strategy for voice recognition. To begin, use the vocal tract length perturbation (VTLP) technique to enhance the data set, and then use noise perturbation technology based on genetic algorithms to enhance the data set again. To accomplish the objective of speech recognition, they proposed the DeepSpeech2 model, and the performance of the model is shown good. Donahue et al. [18] first used GAN to generate sounds data by analyzing sounds data in time and frequency domain. Since the loss function of GAN can only be used as a measure of the similarity between the generated data and the real data distribution. Arjovsky et al. [19] proposed a new GAN variant, Wasserstein GAN, to solve this problem. Based on this study, a new gradient function for Wasserstein GAN loss is introduced in [20]. Recently, based on the theory of WaveGAN, a new kind of WaveGAN network variant parallel WaveGAN is proposed in [21]. This method can effectively capture the time-frequency distribution of the real speech waveform by training a non-autoregressive wave network through joint optimization of multi-resolution spectrum and anti-loss function.

Therefore, based on the theory of traditional sounds data augmentation technology, this paper tries to find a way to generate reliable and near real log Mel spectrum on GAN. DCGAN and CapsuleGAN are tested and analyzed. Because Capsulenet can recognize the relationship between the features of data distribution and the relationship between the features and the whole, combining with Griffin Lim algorithm, an end-to-end fault sound generation system, Mechanical failure sound CapsuleGAN (MFS-CapsuleGAN), based on CapsuleGAN is proposed to augment the data.

2.2 Generative Adversarial Network (GAN)

GAN consists of a generator and discriminator. In GAN, in the beginning, the generator receives random numbers generated from the Gaussian distribution and generates false data after the deconvolution operation of up-sampling. The fake data and the real data are labeled and sent to the discriminator. The discriminator learns real and false data samples and obtains the network model with strong discrimination ability after continuous training. The trained discriminator has a better ability to identify whether the sample is from real data or fake data generated by the generator. In the discriminator network, when the input is real data, the output of the discriminator network is close to 1. When the input is false data, and the output of the discriminator network is close to 0. That proves the discriminator has good performance and achieves the ideal goal of identification. When the performance of the generator is poor, the discriminator can easily identify the fake data. The discriminator feeds back to the generator the difference values between the data generated by the generator and the real data. The generator constantly updates the parameters to optimize the network, so that the generator network learns towards the direction of decreasing the D-value, to produce more real false data.

To effectively distinguish the true data from the false data, the discriminator needs to train itself to improve its discrimination ability through the real data to prevent the generator from mixing the false data with the true data. Through this mutual antagonism, the discriminator and the generator carry on the reverse training to each other. Finally, as the fake data gets closer and closer to the real data, the discriminator can't distinguish whether the data generated by the transmitted generator is from the real data distribution or the false data distribution so that the generator can achieve good performance.

In GAN, the discriminator and generator are respectively composed of convolutional neural network and deconvolution neural network. The operation of traditional CNN forward propagation is to compress the size of feature graph and make it smaller and smaller, while deconvolution is to make the initial input data (noise) bigger and bigger. After several layers of convolution, false data with the same dimension size as the original data is obtained. DCGAN defines a noise $P_z(z)$ as a prior, which is used to learn the probability distribution P_g of the generator network model G on the training data x, and $G(z)$ represents the mapping of the input noise z into the data space. $D(x)$ represents the probability that x comes from the real data distribution P_{data} instead of P_g. We trained D to maximize the probability of correctly distinguishing real samples from generated samples, so we can train G by minimizing $\log(1 - D(G(z)))$ at the same time. In other words, discriminator D and generator G play minimax game on value function $V(G, D)$. Accordingly, the optimized objective function is defined as follows:

$$\begin{aligned}\min_G \max_D V(G,D) &= \mathbb{E}_{x\sim P_{\text{data}}(x)}\left[\log D(x)\right] \\ &+\mathbb{E}_{z\sim P_z(z)}\left[\log\left(1-D\left(G(Z)\right)\right)\right]\end{aligned}. \tag{1}$$

If the parameters of the discriminant model D are updated, then for sample x from real distribution P_{data}, we hope that the output of $D(x)$ is as close to 1 as possible, that is, the greater F is, the better the discriminant ability of D is. For the data $G(z)$ generated by noise z, we want $D(G(z))$ to be as close to 0 as possible, so the larger $\log(1 - D(G(Z)))$ is, the stronger the discrimination ability of D is, so we need to maximize D.

If the parameters of the discriminant model G are updated, then we want $G(z)$ to be as real as possible, which is $P_g = P_{\text{data}}$. Therefore, we want $D(G(z))$ to be as close as possible to 1, that is, the smaller $\log(1 - D(G(Z)))$ is, the better the performance of generator is, so we need to minimize G.

In theory, if the fixed G value updates D, then the optimal solution is:

$$D^*(x) = \frac{P_{\text{data}}(x)}{P_{\text{data}}(x)+P_g(x)}. \tag{2}$$

However, when updating G, the target function takes the global minimum if and only if $P_g = P_{\text{data}}$. Ideally, after constant parameter updates and network optimization, the final result of the game between the two models is that G can generate $G(z)$ that can be regarded as the real data. Therefore, it is difficult for D to judge whether the data generated by G is true, that is, $D(G(z)) = 0.5$.

The principle of GAN is to let D and G play games, and the two models can be simultaneously augmented by competing with each other in the training process. Due to the supervision effect of the discriminant model D, G can produce false data close to the truth without a large amount of prior knowledge and prior distribution.

3 Data augmented method

In Section 3, we focus on the spectrum data augmented method based on MFS-CapsuleGAN, improved CapsuleGAN, and apply it to the characteristic data augmented method for log Mel spectrum of mechanical fault sounds.

3.1 Spectrum Box Selection Data Augmentation (SBSDA)

Usually for processing data, waveform audio is converted into a spectrogram and then fed into a neural network to generate output. The traditional way of performing data augmented is usually applied to waveforms. Seeking another method is to manipulate the spectrogram, based on a priori algorithm in the field of image, they make image to flip, rotate, scale, crop, shift, add Gaussian noise, color augmented and so on. These operations also apply to the spectrogram. Park et al. [22] proposed SpecAugment for data augmented in speech recognition. There are three basic ways of expanding data, namely time warping, frequency masking and time masking:

Time warping: A random point will be selected and bent to the left or right along the distance w, and the distance w will be selected along the line from 0 to the uniform distribution of the time warping parameter W.

Frequency masking: The channel is masked. It is selected from 0 to a uniform distribution of the frequency template parameter F, and is selected from among them, where is the number of frequency channels.

Time masking: t consecutive time steps $[t_0, t_0 + t)$ are masked. The t is selected from the uniform distribution of the time mask parameter T from 0, but t_0 is selected from $[0, \tau - t)$.

Compared with data augmented on sound signal waveform, data augmented on frequency spectrograms can improve training speed, because there is no need to perform data conversion between audio signal data and spectrogram data, while spectrogram data can be increased. However, this method still cannot completely solve the remaining problems by the aforementioned audio signal data augmented.

The log Mel spectrum based on the input audio proposed by Park et al. [22] instead of the augmented method of the original audio itself [23]. The theoretical basis provided is based on the transformation of the spectrum. The theoretical basis provided by this method can compare the spectrum to the image, perform frame-selecting data on the spectrogram, and avoid multiple conversions from audio signals to spectrogram signals, as shown in Fig. 1.

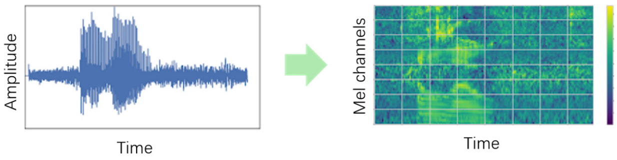


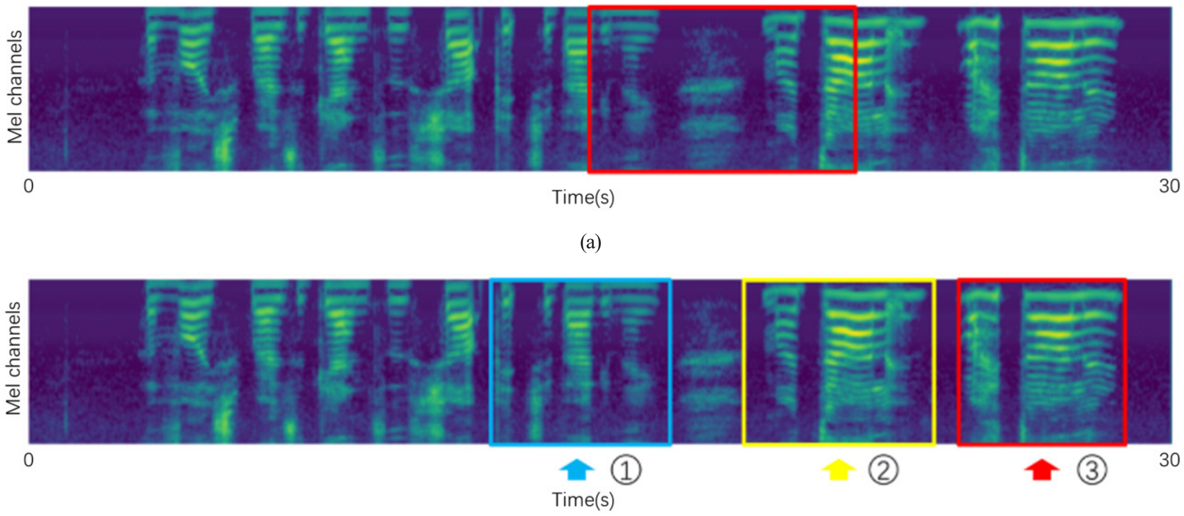
Fig. 1 Schematic diagram of waveform audio to spectrum diagram

The fixed matrix data input settings of CNN and DNN will often miss the feature connection between two adjacent signal frequency spectrums when training CNN with frequency spectrum, which will cause when the neural network to train according to the pre-segmented signal data, the features in the fixed spectrum will be considered only, and the spectral features of missing features or incomplete signals will be misclassified or even not recognized. As shown in Fig. 2 (a), this kind of problem often appears in the training set. The probability of occurrence in the verification data set during actual field detection is greater, resulting in a higher probability of underfitting the neural network in the case of data scarcity.

To solve the above problems, this study proposes a log Mel feature augmented method for feature refinement and enlargement of the unrestricted spectral window size based on the frequency spectrum data augmented method. This method is dedicated to solving single-fault signal features.

As shown in the schematic diagram of the log Mel random frame selection in Fig. 2, we select a continuous stream of fault audio for up to 30 s instead of a 3–5 s pre-segment fault audio segment.

For example, as seen in Fig. 2 (a), if we use the duration of the frame with the length to cut the pre-segmented audio at this position, although it contains the sound features between the complete fault signal and the signal, most of the signal features of the log Mel spectrum of this segment do not contain the fault signal features, only a small part contains the fault signal. The disadvantages of segment features are that there is still a difference in the distribution of two-dimensional feature data of the same type of fault signal. The duration of the fault audio set in advance cannot completely contain all the features of at least a single fault signal, which leads to the learning of the neural network may be an incomplete fault signal feature. On the



(a)

(b)

Fig. 2 Schematic diagram of log Mel random frame selection, (a) the same length of the pre-segmented audio, (b) the smaller segmented audio

contrary, although the smaller segment in Fig. 2 (b) can contain the complete spectrum characteristics of the fault signal, there may also be frequency bands that do not contain the characteristics of the fault signal, and they cannot reflect the spectrum characteristics between signals.

In order to solve the above problems, under the condition that the number of features is the same, that is, the number of rows of the spectrogram matrix is the same, the frame length in the region of the selection is randomly and not repeatedly. We set the total frame length F of the spectrogram, and the distance between the two fault signals within the range of $[N, M]$. The maximum period of the signal, which can be regarded as the frame length within the range of $[T1, T2]$, and the randomly framed spectrum frame length is F', then its definition domain is: $[N + 2T1, M + 2T2]$. Among the selected spectrogram features, there is at least one complete fault signal spectrogram feature and non-fault signal feature including at least one with an approximate maximum probability. The schematic diagram on the time domain diagram can be expressed as shown in Fig. 3.

In addition, in order to refine the features, the minimum frame number $F_{\min}$ is set as the minimum selectable frame number, the step size is L, and the randomly framed spectrum frame length is $(F_{\min}, F - F_0)$. It will maximize the creation of a frequency spectrum in an L step, as much as possible while increasing the number of features, but also play a role in refining features.

3.2 MFS-DCGAN for data augmentation

The main component of MFS-DCGAN is the generator and discriminator. The generator consists mainly of three deconvolution layers.

The purpose of the generator is to map from the potential space to a specific distribution close to the real data distribution. The network architecture of the generator is described in detail below.

The first layer is the input layer, which receives the noise sample data. Take the size of the HAASD sample data for example. The noisy sample z is the Gaussian distribution data generated randomly from −1 to 1 with dimensions of 64 × 432, then the data was batch normalized.

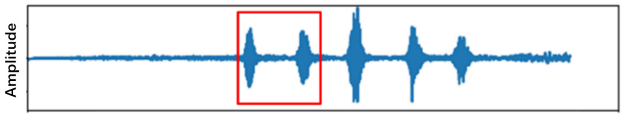


Fig. 3 Schematic diagram of random frame selection

The second layer is the deconvolution layer, consisting of 256 kernels with each size of 5 × 5, and we stride the filter by 5 × 5. The pooling layer is eliminated in the deconvolution layer, the pooling method is replaced by deconvolution method to up-sampling for the sample.

The third layer is also deconvolution one, whose structure is consistent with the second layer. The input of this layer is the output of the second layer, and 128 convolution kernels are used for deconvolution. After batch normalization and activation, a dropout function with a probability of 0.5 is added to the output to prevent over-fitting.

The fourth layer is the last layer of the deconvolution layers, where the input is the output of the third layer. The parameters of kernel and stride are consistent with the second and third layers, and the deconvolution result will be output after the activated by Leaky ReLU function.

The purpose of the discriminator is to try to distinguish between the data fed to it, which is the real data distribution, and which is the fake data distribution generated by the generator. Here is the detailed architecture of the discriminator.

The first layer is the convolution layer, with the real data and the data generated by the generator as input. 128 convolution kernels with each size of 4 × 4 were used to downsample the feature map, and the stride size of convolution is 4 × 4. After that, Leaky ReLU activation function will be used to activate this layer before output.

Both the second and third layers are convolution layers, the second layer uses 256 kernels, while the third layer uses 512 kernels, and the other architectures are consistent with the first layer. The difference is that batch normalization should be carried out after the convolution in the second layer, and the dropout with a probability coefficient of 0.5 should be added after the activation function in the third layer to prevent over-fitting.

The fourth layer is the output layer, and the sigmoid function is used for the probability distribution. The maximum probability of the final output represents whether the input into the discriminator is from real data or fake data generated by the generator. More details of MFS-DCGAN architectures are shown in Fig. 4.

3.3 MFS-CapsuleGAN for data augmentation

MFS-DCGAN, which is based on convolution and deconvolution networks, relies on the characteristics of CNN architecture, resulting in the omission of some data. The convoluted feature spectrum may not contain some feature information of the previous layer. The decrease

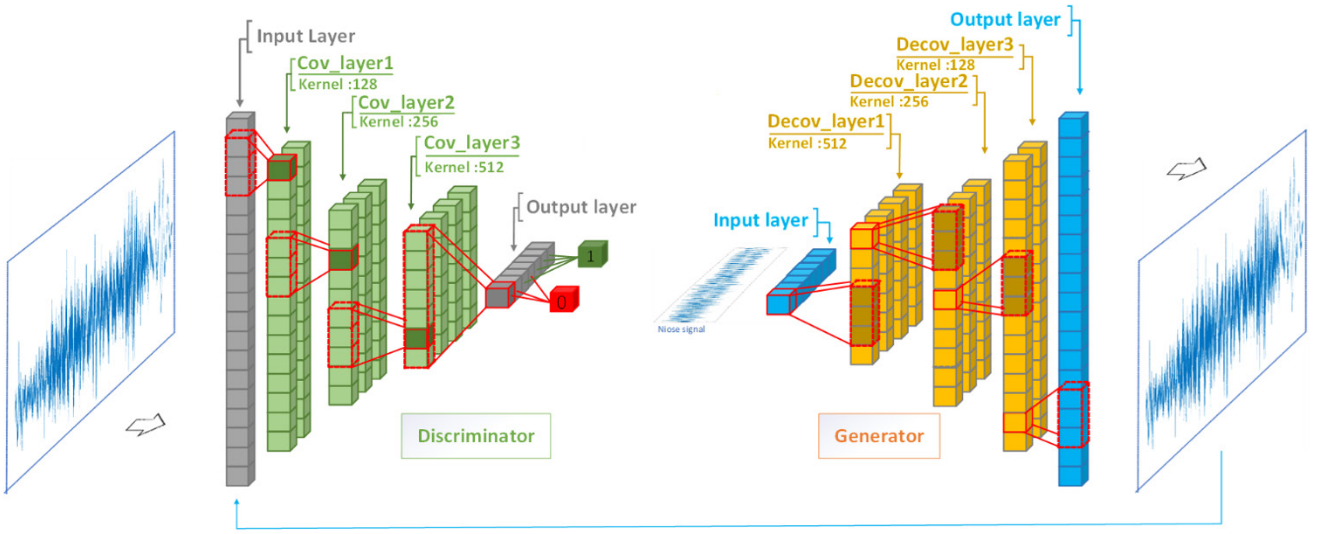


Fig. 4 MFS-DCGAN network diagram: 1 and 0 are tags of normal and abnormal sound, respectively, indicating that the discriminator can distinguish the input sound samples from real data or fake data generated by MFS-DCGAN

in visual field information extraction ability leads to a decrease in spatial resolution, which cannot deal with fine features well. The result of the characteristic spectrum for small changes in input is often almost constant. It is not sensitive to the position relation between features, but the characteristic relation between frames in the spectrum. The problem can now be solved by building capsule neurons, training the data in vector form, and iteratively updating the parameters through routing algorithms. In CapsuleNet, the position information of spectral features, rotation, thickness, inclination, and size of feature representation is learned by the capsule network as images, so that the pooling operation will not be lost first and then recovered. The small changes of low frequency and weak signals generated by mechanical vibration can also be learned by the network, which is called "equivariant". That is, CapsuleNet can use a simple and unified architecture to handle the task of feature spectrum recognition. The discriminator of MFS-CapsuleGAN is similar to the CapsuleNet model in structure. In general, CapsuleNet has a large number of parameters. Because first, each capsule generates a vector output rather than a single scalar, and second, each capsule has additional parameters associated with all the capsules in the layer above it to predict its output. We use the marginal loss LM instead of the traditional binary cross entropy loss to train our MFS-CapsuleGAN because LM is more suitable for training the CapsuleNet discriminator.

MFS-CapsuleGAN composed of CapsuleNet, which can get feature connection, has better feature generation ability than DCGAN constructed by CNN. CNN architecture can't analyze feature location relationships well, while CapsuleNet can also perform well in complicated feature data. To assess the generation effect of the MFS-CapsuleGAN model, an end-to-end fault audio spectrum data augmented experimental system is built. The system obtains the characteristic data of the log Mel spectrum from the audio signal through preprocessing. Based on these original sample data, the random spectrum box selection algorithm is used to enhance the data and obtain the original augmented data without changing the characteristics. These original data augmentation sets are augmented by MFS-CapsuleGAN, and CGAD (CapsuleGAN augmented data) is generated. To visually verify the authenticity of CGAD, the fast Griffin-Lim algorithm [24] is used to reconstruct the audio signal of randomly selected data when MFS-CapsuleGAN data is output. When each batch of data is augmented, the randomly generated audio signal sample data is output for the human ear and spectrum analysis verification. The system is shown in Fig. 5.

The MFS-CapsuleGAN generated confrontation network model based on spectral graph characteristics as input and output are similar to image recognition. In this paper, single-channel data training is adopted, and the input and output are all two-dimensional matrix data of spectral graphs. The network consists of a generator network composed of a deconvolution layer and a discriminator composed of a capsule neural network. Fig. 6 shows the capsule generation antagonistic network model.

The network consists of a deconvolution generator and a Capsule discriminator. Taking washing machine sample data as an example, the input is two-dimensional matrix,

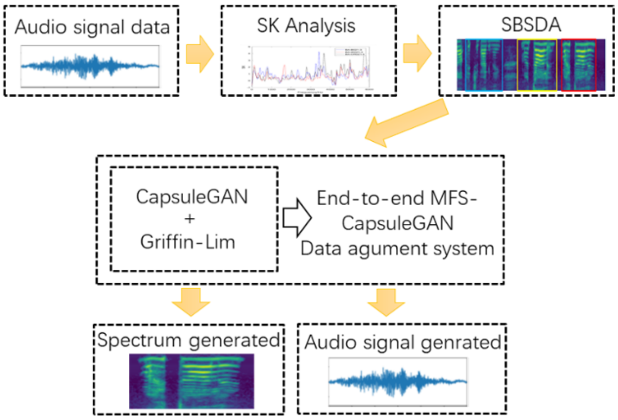


Fig. 5 End-to-end log Mel spectrum data capsule generation countermeasure augmentation system

and the output is $n \times 1$ probability vector, where n is the fault type. The first layer of the discriminator is the convolutional layer. Compared with the most advanced convolutional neural network variants in the industry, it is a very shallow network with only four convolutional layers and a fully connected layer in the middle. 256 convolution Windows of size 9 × 9 are used, and get an output of 20 × 20 × 256. 32 filters with size of 9 × 9 × 256, stride is 2, be used in primary capsule layer to obtain the output with dimension of 6 × 6 × 32. This is equivalent to eight convolution operations with 32 filters of step 2. And in Capsule layer, each element in the dimension of 6 × 6 × 8 × 32 is a 1 × 8 vector. The next layer is to store vectors of high-level features. The PrimaryCaps layer and DigitCaps layer are fully connected, but unlike the traditional neural network scalar and scalar, they are connected with vectors, and finally output the maximum probability of discrimination. MFS-CapsuleGAN has a certain robustness, and the model obtains losses by reconstructing the difference between the log Mel image and the real image. In this paper, the reconstruction loss is the subtraction and square summation of the pixel values on the 6,487 neural units of the final output and the initial input, which is defined as:

$$T_{\text{loss}} = G_{\text{loss}} + \alpha \times M_{\text{loss}} \,. \tag{3}$$

In the above $\alpha = 0.01$, is the reconstruction loss, T_{loss} is the interval loss, and M_{loss} occupies a dominant position.

4 GCMD network

The GCMD network can be regarded as a variant of the traditional artificial neural network structure. It no longer uses the fully connected hidden layer but consists of the convolutional layer, the pooling layer, and the fully connected layer [25].

We used the structure of a convolution neural network for model training. This network uses four layers of convolution and two layers of full connectivity.

Take 5 s long fault sample training as an example. The first layer is the convolution layer, the size of the input single sample is 64 × 432, and this convolutional layer has 32 filters with the size of 5 × 5 we stride the filter by 1. After the convolution, it was activated by the Leaky ReLU function, and then the outputs data were normalized by batch normalization. The traditional neural network only

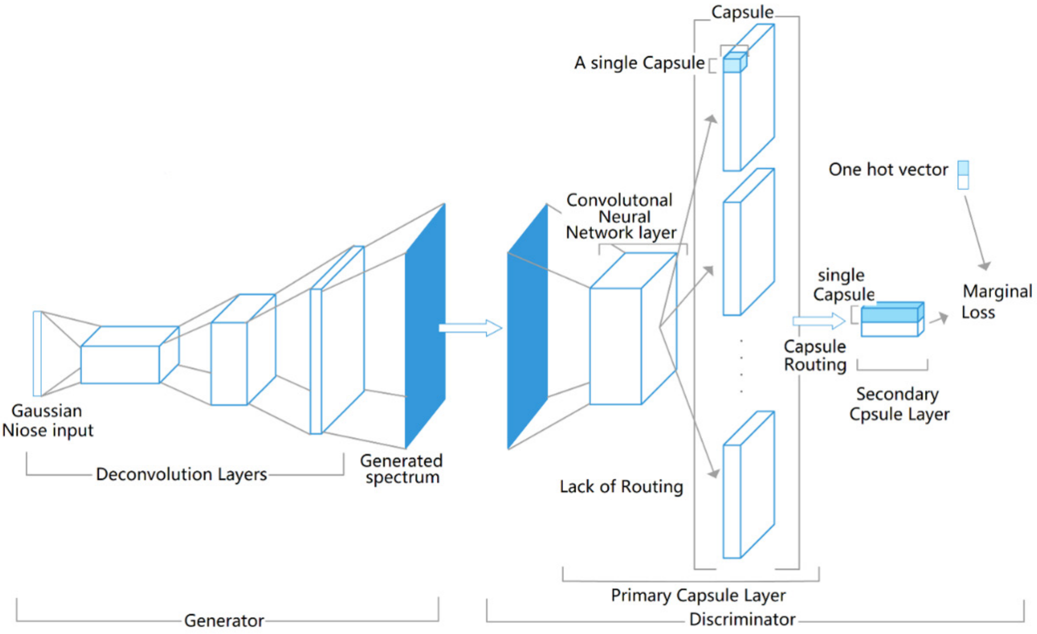


Fig. 6 MFS-CapsuleGAN model

conducts data normalization processing before inputting sample data was fed into the input layer to reduce the difference between samples. However, on this basis, batch standardization not only standardizes the input data of the input layer, but also standardizes the data of each hidden layer to achieve the purpose of preventing gradient diffusion and accelerating convergence. After standardized processing, the data will be fed to pooling layer. The pooling window size was 5 × 5 with stride of 5 × 5 and it was selected to output 32 feature maps.

The second layer is the convolution layer. The output of the second layer is used as input for the third layer, and 64 filters are used for convolution with a size of 3 × 3 and we choose to stride the filter by 3 × 3. It was first activated by Leaky ReLU activation function, then normalized, and finally pooled by pooling layer with a pool size of 2 × 2, stride size of 2 × 2, and output 64 feature maps.

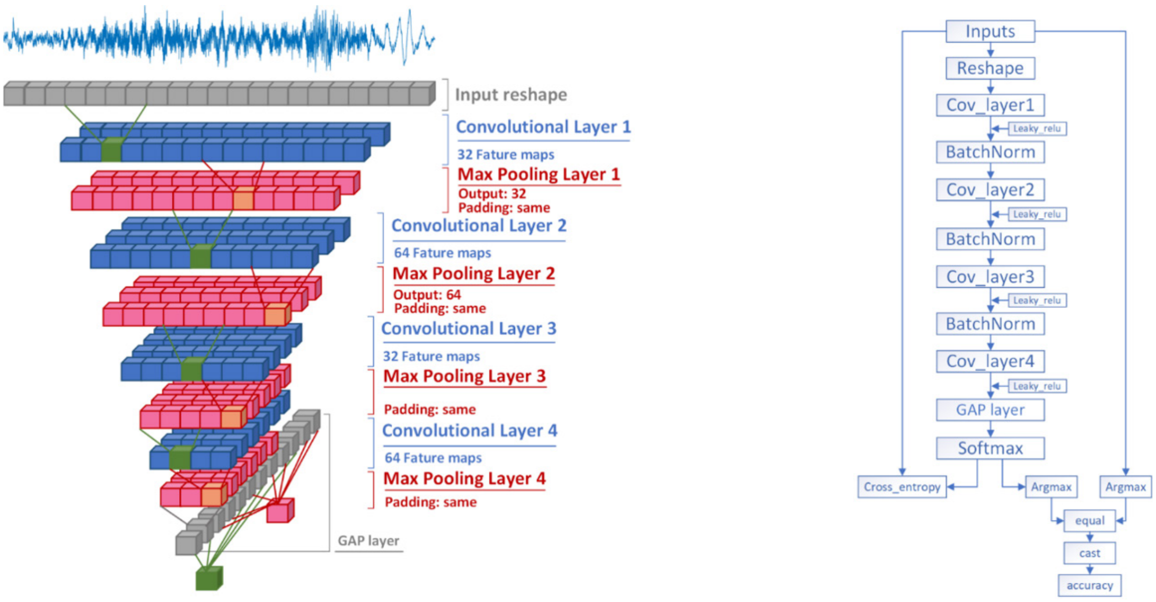
The third and fourth layers are also convolution layers, and the output of the second and third layers are respectively used as the input of those two layers. Activation and batch normalization parameters are consistent with layer 2. Convolution kernel size, strides, pooled window size and the padding method are all the same as the second layer. The difference is that the third layer uses 128 kernels for convolution, and outputs 128 feature maps, the fourth layer uses 256 kernels and outputs 256 feature maps.

After last convolutional operations, outputs will be fed to the GAP layer. Inputs will be reshaped to one-dimension, and multiply it by the weight, add the bias, then activate it through Leaky ReLU activation function. To overcome over-fitting problem, we use dropout regularization technique to fix it. At the end of fully connected layer. Dropout discards certain neurons in the fully connected layer with a probability of *p*. After that, it is then fed into the output layer. In output layer, inputs of this layer will be mapped into *n* output class, and the Softmax function is used to assign probabilities to each of these samples. More details of GCMD architectures are shown in Fig. 7.

The traditional convolutional neural network performs convolution on the hidden layer of the network. For classification, the feature map of the last convolutional layer is vectorized and input into the fully connected layer and finally passes through the Softmax or logistic regression layer. This structure connects the convolution structure with the traditional neural network classifier. It uses the convolutional layer as a feature extractor and classifies the obtained features using traditional methods.

However, the fully connected layer is prone to overfitting, which affects the generalization ability of the entire network. Dropout was proposed by Gomez et al. [26]. As a regularizer, half of the activations of fully connected layers are randomly set to zero during training. It improves the generalization ability and largely prevents overfitting.

In this paper, the CNN convolution layer is added to the global average pool layer in the experiment to replace the traditional fully connected layer. The idea is to generate a feature map for each corresponding category of the classification task in the last convolutional layer. We do



(a)

(b)

Fig. 7 CNN model architecture for Abnormal Sound Detection of Machines. (a) The detailed composition of each CNN layer structure, (b) CNN's tensor flow diagram and the modes of activation added to each layer

not need to add a fully connected layer on top of the feature map, but take the average of each feature map, and the resulting vector is directly input to the Softmax layer to convert from a high-dimensional space. As shown in Fig. 7, we suppose that the final output of the convolutional layer $h \times w\times d$ is a three-dimensional feature map and the specific size is 6 × 6× 3. After GAP conversion, it becomes an output value of size 1 × 1× 3, that is, each layer will be averaged to a value $h \times w$.

As shown in Fig. 8. Compared with the fully connected layer, an advantage of the global average pool is that it is more suitable for convolutional structures by enhancing the correspondence between feature maps and categories. Therefore, feature maps can be easily interpreted as category confidence maps. Another advantage is that there are no parameters to optimize in the global average pool, so overfitting is avoided at this level. In addition, the global average pool merges the feature space information to make the input space translation more robust. You can think of the global average pool as a structural regularizer, forcing feature maps to be confidence maps of concepts (categories). This is achieved by multiple convolutional layers because they are easier to classify than GLMs. In short, after the convolutional layer, replacing the FC fully connected layer with GAP. There are two advantages: one is that GAP is simpler and more natural to convert between the feature map and the final classification; the second is that, unlike the FC layer that requires a lot of training and tuning parameters, reducing the spatial parameters will make the model more robust and resist overfitting effects better.

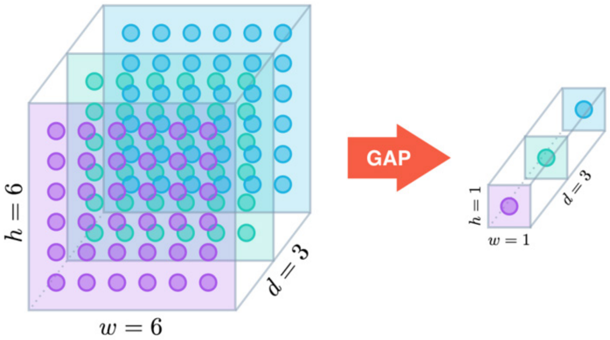


Fig. 8 Schematic diagram of GAP conversion

As shown in Fig. 9, the last fully connected layer is removed and replaced by the GAP layer, which extracts the final feature map information in the form of a fixed output, and each feature map outputs a value. In this way, when the size of the convolution window is allowed, no matter what the size of the input log Mel spectrum window is, there is no need to continue to adjust the convolution parameters, which solves the inconvenience of the time limit for the actual detection on site. It also provides a theoretical basis for the feasibility of the design of random frame selection data augmentation algorithms.

5 Experiments

Firstly, SBSDA was used for the GCMD network training experiment, and then CGAD was used for MFS-DCGAN and MFS-CapsuleGAN training experiment and its data generation effect analysis. Finally, a comparative analysis of the fault speech recognition network model is carried out to verify the performance of GCMD and its robustness.

To effectively and efficiently verify the fault source of sound source characteristics, the fault characteristics are highlighted by spectral envelope analysis after noise reduction filtering. As we can see from Fig. 10, the washing machine drum failure was obtained in a frequency range of 250 Hz, where the signal showed a higher modulation effect. The basic frequency is 76.02 Hz, and the second and third harmonics are 146 Hz and 211 Hz. The amplitude of the fault frequency and its harmonics are clear, in which case the second harmonic offset is small. In the spectral envelope signals of ToyADMOS and MIMII datasets, the amplitude of external fault frequency and its harmonics are clearly visible. Compared with the MIMII datasets, the harmonic frequencies of 200, 400, 600, and 800 Hz are more accurate.

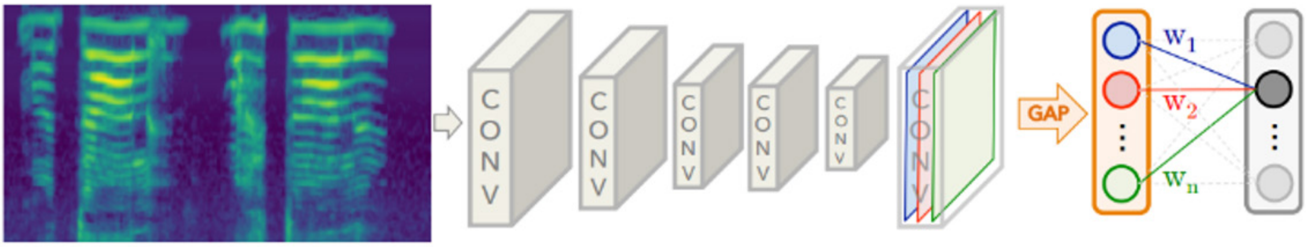


Fig. 9 Schematic diagram of GCMD

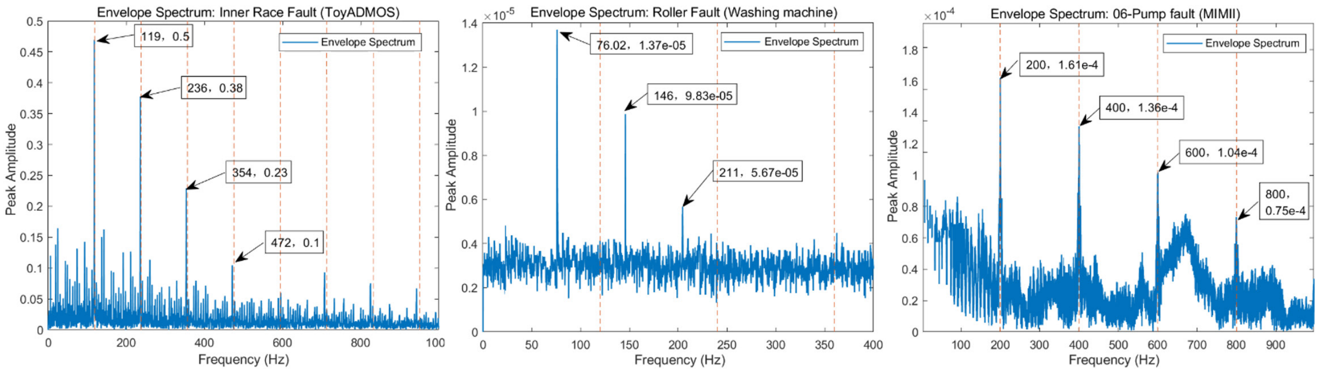


Fig. 10 Spectrum envelope analysis diagram of fault signal

5.1 Data preparation

Log Mel spectrum graph was selected in a random box, and the data was augmented based on 1000 different fault audio signal data samples, and the training set was extracted in the form of the spectrum data cache. During data extraction, wav audio files that have been filtered and validated by envelope analysis after fault signals are used as datasets, and each sample lasts for the 30 s. After feature extraction, the box selection algorithm is used for data augmentation. Each window of the random box was selected as the new spectrum size, and the augmented data of different sizes were used as a batch of data. 50 random points were set in the experiment, and the random value was selected as an integer value, without repeated random selection within the optional range of frame *F*.

5.2 MFS-CapsuleGAN

We use the Gradient Descent Optimization to replace the Adam Optimization in MFS-CapsuleGAN. It was found that Gradient Descent Optimization converges better than Adam Optimizer in the optimization process. Cross entropy is used to define generator loss and discriminator loss. Based on MFS-CapsuleGAN, we added the smoothing parameter of labels, to alleviate the problem that the label is not soft enough and can easily lead to overfitting so that the model is less confident in predicting. Through tests, it was determined that smooth = 0.25 and the generator and discriminator performed best. The smoothing parameter is added so that the discriminator is not overconfident, and the generator loss value is much larger than the discriminator.

We set the parameters and get the loss of the current generator and discriminator once every 100 iterations of training. As shown in Fig. 11, as the times of training increase, changing the smooth value will affect the *D*-value between the loss value of the discriminator and the generator.

One notable drawback of MFS-CapsuleGAN is that it is prone to gradient explosions. In network training, batch normalization can be added to the layer to solve the problem of slow convergence speed or gradient explosion. In addition, batch normalization can also be added to speed up training and improve model accuracy. By comparison,

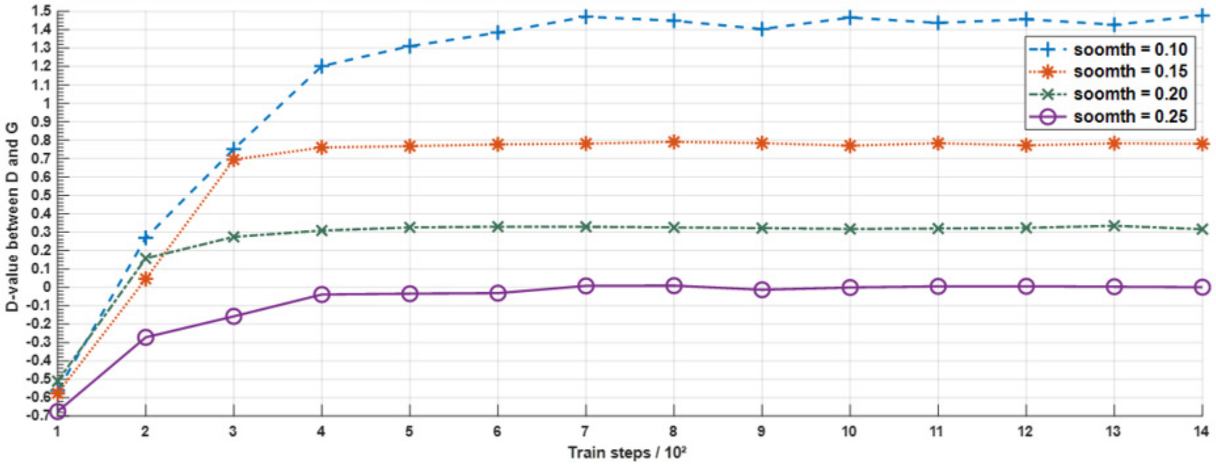


Fig. 11 The difference between the loss values of the generator and the discriminator network

Leaky ReLU activation function performed better than Tanh activation function or ReLU activation function. Since most negative values exist in the original data, if Tanh or ReLU activation function is adopted, this part of value will be weakened, while Leaky ReLU can solve this problem. It can correct the data distribution and retain valid values, so that all negative values will not be lost. Therefore, we used the Leaky ReLU activation function to activate each layer before output.

The experiment was conducted for 140,000 iterations of training. The networks are trained on a workstation equipped with an E5 6230 CPU with 32 GB RAM and 4 NVIDIA Tesla K40t GPUs with 12 GB RAM of each.

Training parameters and experimental equipment for MFS-CapsuleGAN in Table 1.

Advantages of adding label smoothing:

1. First, to ensure the generalization ability of the model and prevent over-fitting.
2. Secondly, it can be known from the gradient boundness that minimizing the total probability and the zero probability encourages the difference between the category and other categories, which can effectively prevent the phenomenon that the loss value of generator is much larger than that of discriminator caused by the model's over-belief in the predicted category.

Table 1 Training parameters and experimental equipment

Learning rate	$\eta = 0.0001$
Dropout probability	$p = 0.5$
Smooth	0.25
Batch size	1000
Epochs	140000
CPU	E5 6230
RAM	32 GB
GPU	Tesla K40t

We set the parameters and get the loss of the current generator and discriminator once every 30 iterations of training. The loss value of the MFS-DCGAN model is shown in Fig. 12. It can be seen from Fig. 12 that when the loss value is between 0.1 and 0.2, the loss value of generator and discriminator is greatly different. While the value is set as 0.25, the difference is very small, close to 0. After training, the output of the generator and discriminator are relatively stable, and the loss value is close to 1. At this point, the discriminator cannot distinguish the real sample from the fake sample.

The loss value of the MFS-CapsuleGAN model is shown in Fig. 13. We can see that after 150,000 iterations, the generator and discriminator losses are close to zero, and the network tends to stabilize. It can be seen that the loss of the generator and discriminator increases negatively at the beginning of training. With the increase in the number of iterations, the loss value of both approaches to 0, which is close to the ideal state, indicating that the training process is very smooth, and the performance of the generator and discriminator is close to the balance.

To visualize what happens to the spectrum. We will produce part of the spectrum from reshape as 64 × 64 dimension, to observe the diversity of the generated data change. As we see from the spectrum diagram in Fig. 14, each row represents 6 kinds of varieties of a fault spectrogram.

In order to verify the improvement of model performance by data augmentation. In the training experiment of GCMD network, the cross-entropy function is used to calculate the loss in the training network. The Adam optimizer, whose vertical learning rate η is 0.0001, is used to update the parameters (w, b), and the Dropout probability p is set as 0.5. Although the use of Dropout resulted in increased training time, it improved the performance of neural networks in supervising learning tasks. The equal

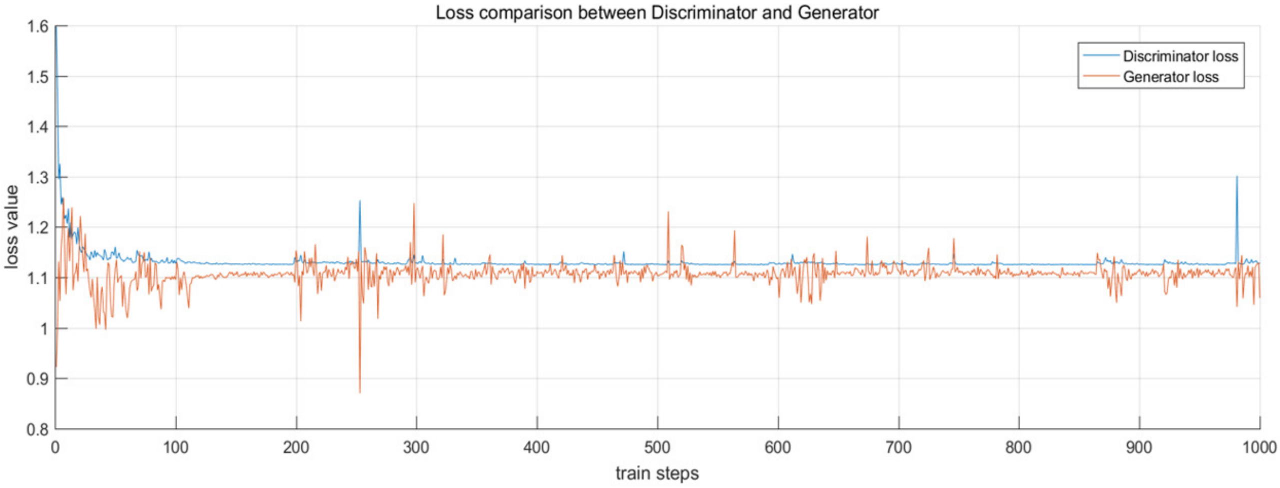


Fig. 12 Different loss values for the discriminator and the generator

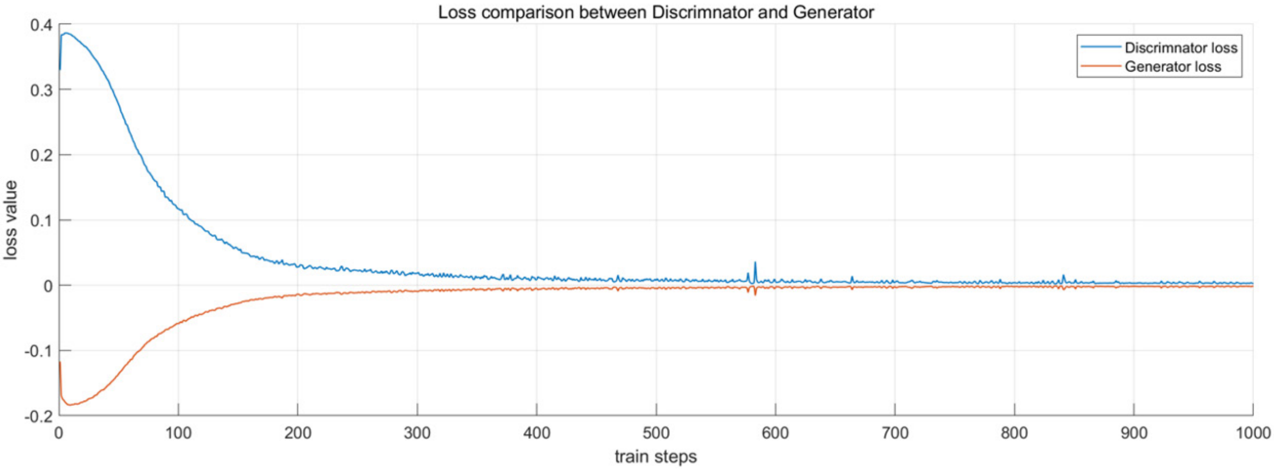


Fig. 13 Loss values of the discriminator and generator approach equilibrium

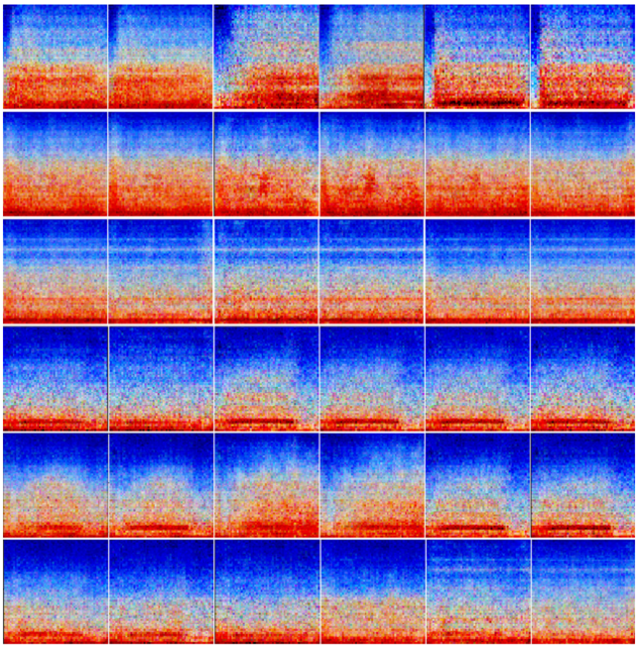


Fig. 14 Spectrum generated for 6 machine faults

function compares the prediction value with the true value, returns true or false, is converted into a 1 or 0 format using the cast function, and then averages the calculation, thereby calculating the accuracy. GCMD network is trained using a batch size of 1000 with 400,000 epochs in this work. Each iteration outputs training accuracy, and finally tests and then outputs test accuracy. CNN classifier is trained on a workstation equipped with an AMD r7-3900X CPU with 64 GB RAM and 4 RTX 2080TI GPU with 11 GB RAM.

Training parameters and experimental equipment for GCMD in Table 2.

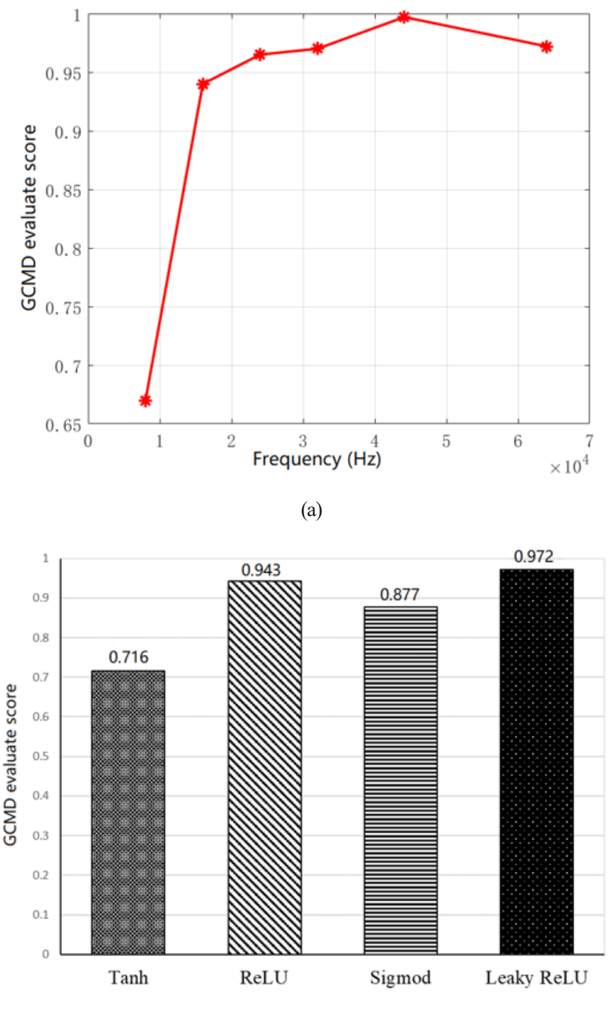
Network parameter tuning is carried out during network training. Digital audio signals can be collected at different sampling rates. However, it is not certain which sampling rate is most suitable for the acquisition of fault sound, so the experiment carries out an empirical study on different sampling rates and classification accuracy. In the experiment, the idling sound of the washing machine was sampled at the usual sampling frequencies of 8000 Hz, 11025 Hz, 16000 Hz, 32000 Hz, 44100 Hz, and 64000 Hz. In this experiment, the machine idling sound was collected at different sampling rates on the production line, and then the GCMD network was trained with this data set.

Table 2 Training parameters and experimental equipment

Learning rate	$\eta = 0.0001$
Dropout probability	$p = 0.5$
Batch size	1000
Epochs	400000
CPU	AMD r7-3900X
RAM	8 GB
GPU	RTX 2080TI

Fig. 15 (a) shows the average classification test accuracy of the augmented data set obtained from the original data set after training and testing of the laundry fault tone network under different sampling rates, which are 0.673, 0.946, 0.963, 0.971, 0.987, and 0.972, respectively. As can be seen from Fig. 15, the data set produced at the sampling frequency of 44.1 KHz is more suitable for the training of the GCMD network.

Different activation functions have different effects on different data and different networks. In order to determine the optimal activation function among several commonly used activation functions, the augmented data set was used in the experiment, and Tanh, ReLU, Sigmoid and Leaky ReLU activation functions were respectively used for training in the convolutional layer of the washing machine's heterogeneous sound recognition network. As the voice data of washing machine contains most



(a)

(b)

Fig. 15 (a) Accuracy of GCMD model after training with datasets of different sampling frequencies, (b) Influence of different activation functions on GCMD network

negative data, the network test accuracy using 4 activation functions is 0.716, 0.943, 0.877 and 0.972 respectively through comparative test analysis. The network tests using the Leaky ReLU activation function were the most accurate. The average network test accuracy of the four activation functions is shown in Fig. 15 (b).

To verify the promotion effect of the augmented data set on the training network, the experiment used the GCMD network to conduct 400,000 iterations of training. The confusion matrix is used to verify the training results of the augmented dataset for GCMD. The classification accuracy of the obfuscation matrix is shown in Fig. 16. The data set of washing machine fault sound is augmented by SBSDA, DCGAN, and MFS CapsuleGAN. The scores of GCMD were 0.764, 0.835, and 0.987, respectively. Compared with SBSDA, the accuracy of the GCMD model of CGAD training was improved by 0.223. The MFS CapsuleGAN data augmentation method is significant in the washing machine failure sound data set, and the accuracy of the single batch data set test and evaluation of the model is close to 0.99.

In this paper, under the same training conditions, the AUC precision evaluation method is adopted to evaluate the augmented datasets of SBSDA and MFS-CapsuleGAN methods of ToyADMOS datasets and MIMII datasets. The classification effect of the GCMD network model is shown in Tables 3–6.

Evaluation and validation of GCMD network using SBSDA and MFS CapsuleGAN augmented dataset. AUC value is a probability value. When a positive sample and a negative sample are randomly selected, the probability that the current classification algorithm ranks the positive sample in front of the negative sample according to the calculated score value is the AUC value. The larger the AUC value is, the more likely the current classification algorithm is to rank the positive samples in front of the negative ones to better classify them.

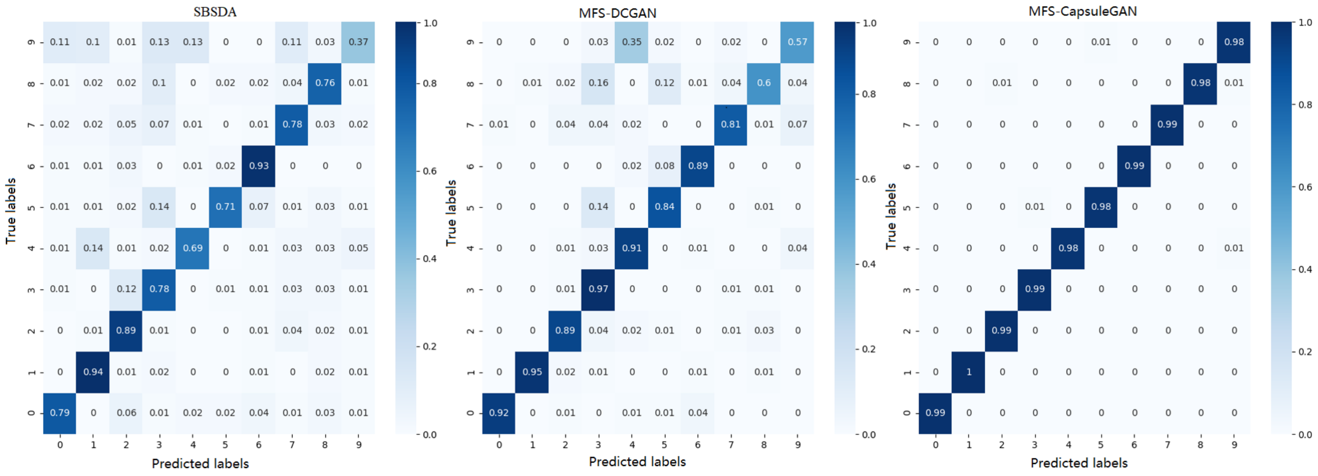


Fig. 16 GCMD evaluates validation accuracy (augmented dataset)

Table 3 ToyADMOS: The AUC and pAUC on the evaluated dataset (SBSDA augmented)

Machine ID	AUC (Ave.)	AUC (Std.)	pAUC (Ave.)	pAUC (Std.)
1	81.46%	1.03%	66.30%	0.94%
2	84.67%	0.59%	76.72%	0.92%
3	66.60%	1.09%	56.32%	0.38%
4	84.93%	1.88%	68.86%	2.44%
Average	79.42%	1.15%	67.05%	1.17%

Table 4 ToyADMOS: The AUC and pAUC on the evaluated dataset (MFS-CapsuleGAN augmented)

Machine ID	AUC (Ave.)	AUC (Std.)	Pauc (Ave.)	pAUC (Std.)
1	89.93%	1.05%	65.15%	1.02%
2	92.22%	1.77%	79.72%	1.41%
3	94.30%	1.02%	86.21%	0.97%
4	90.45%	0.99%	98.97%	2.33%
Average	91.73%	1.21%	82.51%	1.43%

Table 5 MIMII: The AUC and pAUC on the evaluated dataset (SBSDA augmented)

Machine ID	AUC (Ave.)	AUC (Std.)	pAUC (Ave.)	pAUC (Std.)
0	55.31%	0.56%	65.33%	1.02%
2	72.30%	0.65%	58.81%	0.44%
4	60.98%	1.01%	60.33%	0.61%
6	72.21%	0.62%	52.47%	0.43%
Average	65.20%	0.71%	59.24%	0.63%

Table 6 MIMII: The AUC and pAUC on the evaluated dataset (MFS-CapsuleGAN augmented)

Machine ID	AUC (Ave.)	AUC (Std.)	pAUC (Ave.)	pAUC (Std.)
0	90.24%	0.53%	84.44%	1.72%
2	83.43%	0.29%	70.30%	0.84%
4	94.55%	0.62%	76.89%	0.67%
6	87.59%	1.44%	99.01%	1.03%
Average	88.95%	0.72%	82.66%	1.07%

This task is evaluated with the area under the receiver operating characteristic (ROC) curve (AUC) and the partial-AUC (pAUC). The pAUC is an AUC calculated from a portion of the ROC curve over the pre-specified range of interest. In our metric, the pAUC is calculated as the AUC over a low false-positive rate (FPR).

From the experimental results in Tables 3–6, it is not difficult to see that for the ToyADMOS dataset, the recognition accuracy of the MFS CapsuleGAN method is 0.123 higher than that of SBSDA in model performance, and that of the MIMII data set is 0.237 higher.

For the generator generation effect of MFS-CapsuleGAN, it is a problem to verify whether the sample data generated by data augmentation is qualified. Sound is not as intuitive as the image and video. In the actual analysis, the waveform and spectrum can be shown. It is difficult to detect the small change in that waveform or the spectrum by our vision. In addition to subjective analysis of human hearing, this paper builds an end-to-end CapsuleGAN data augmentation system, synthesizes waveform signals through the inverse estimation algorithm of audio signals, and judges the existence of fault signals through spectral envelope analysis. Compared with the intuition of the image, the augmented data generated in the form of log Mel spectrogram is difficult to visually verify the proximity between the data and the spectral features of the real fault signal during the verification, as shown in Fig. 17.

At the same time, through comparative analysis with the original data, the similarity of envelope spectrum failure frequency was analyzed intuitively, and the generation effect of MFS-CapsuleGAN was judged subjectively, as shown in Fig. 18.

In addition, this paper proposes an indirect detection method to test the similarity between the false data generated by MFS-CapsuleGAN and the real data. To solve this problem, we tested the similarity between fake data and real data by indirect detection. The normal or abnormal sound data of the washing machine-generated by MFS-CapsuleGAN is made into a training set, while the original data is used as the testing set. Using this data set to train the CNN network, the average test accuracy was 0.83. This indicates that the fake sample data generated by MFS-CapsuleGAN has obtained most features of the real sample data, and is very similar to the real sample to a large extent, but somewhat different. These differences are just a reflection of the diversity of data, with unknown samples.

The effectiveness of the MFS-CapsuleGANs audio data augmentation proposed in this paper is verified by the data augmentation experiment of the generation countermeasure network. Through the analysis of the experimental results, MFS-CapsuleGAN can improve the generalization ability of the GCMD model much more than the original data set.

5.3 Analysis and results

To verify the superiority of the selected convolutional neural network model in the classification performance of abnormal sound data of washing machines, we compared several representative neural network classifiers to confirm the classification performance of the convolutional neural network.

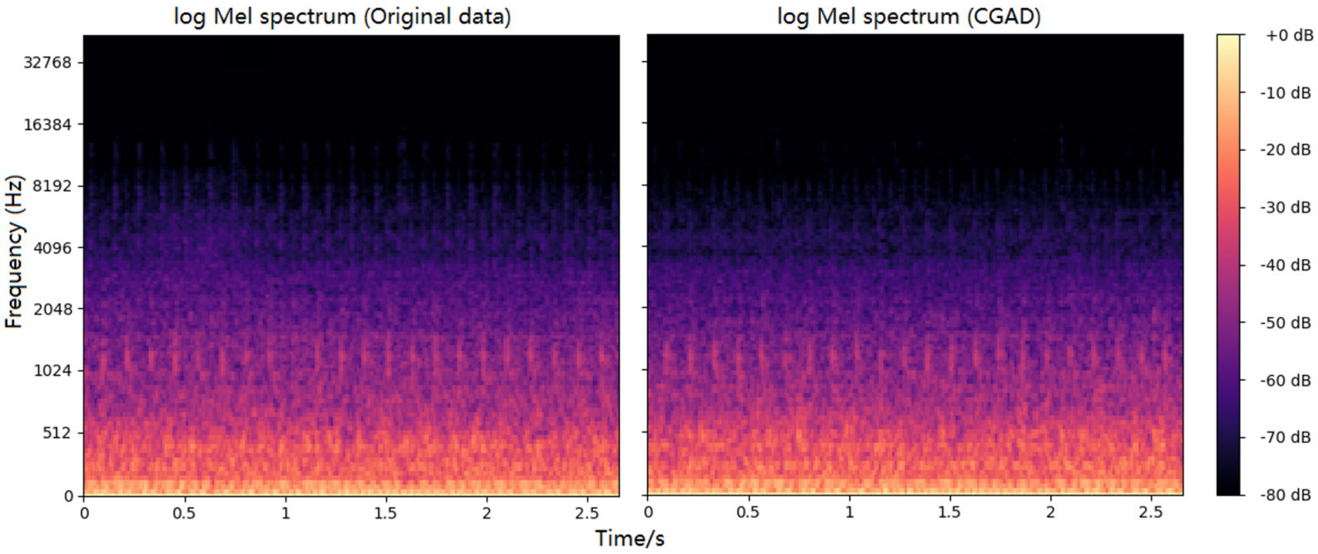


Fig. 17 Comparison of log Mel spectrum between the original data and the generated data

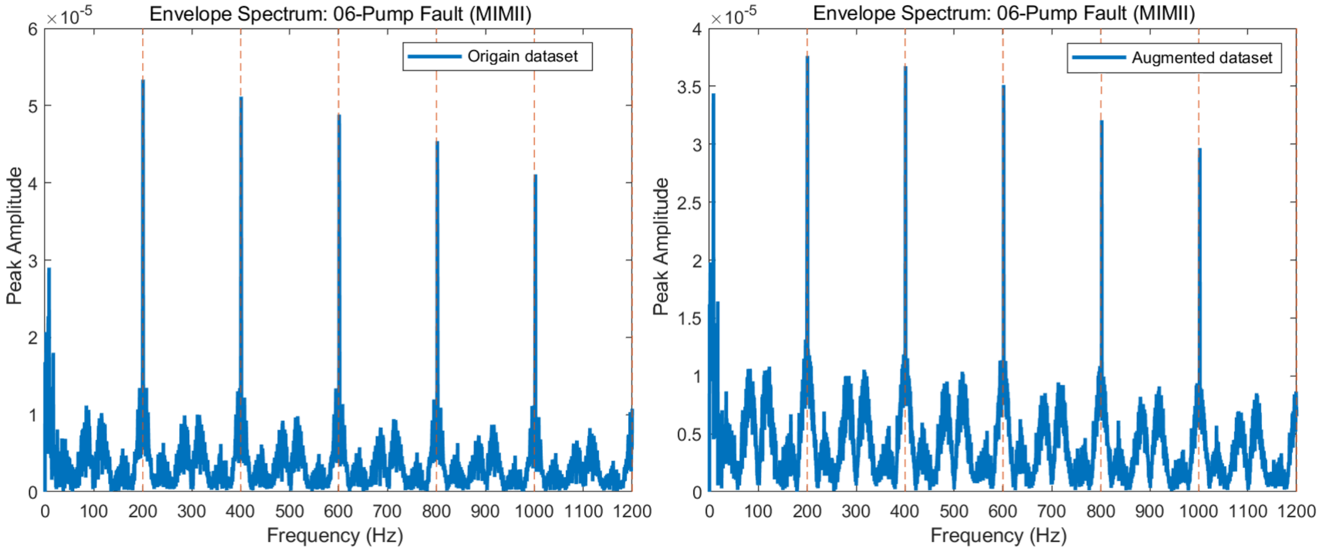


Fig. 18 Comparative analysis of envelope spectrum between the original data and the generated data

To obtain a relatively accurate classification accuracy, we adopted a ten-fold cross-validation method. All datasets were divided into 10 small blocks, 9 of which were randomly selected as training sets and the remaining small blocks as test sets. After ten times training and testing for GCMD, the average test accuracy is obtained, so that the model can obtain the best model parameters.

Six models were trained using the augmented dataset: GCMD, RNN, DNN, SVM, LSTM, and MLP. All six models are trained with the ten-fold cross-validation method. The purpose of using this method for verification is to reduce the contingency caused by the single division of the training set and the verification set. We make full use of the existing data sets to make multiple partitions, to avoid the selection of accidental hyperparameters and models that do not have generalization ability due to special partitions. Therefore, we use cross-validation to reduce contingency and improve generalization ability. We divide all the data into ten copies, and then use each copy as a validation set and the other as a training set for training and validation. In this process, the hyperparameters are kept consistent, and then the average training loss and average verification loss of 10 models are taken to measure the quality of the hyperparameters. Finally, after obtaining a satisfactory hyperparameter, all the data are used as the training set, and the model is obtained by training with the hyperparameter. The results of test accuracy are shown in Table 7. As can be seen, among all trained neural networks, the classification performance of GCMD is the best, and the classification accuracy of deep neural networks is above

Table 7 Cross-validation accuracy of each models

Models	Accuracy
GCMD	0.989
RNN	0.687
DNN	0.992
SVM	0.681
LSTM	0.995
MLP	0.994

0.9. RNN and SVM have the worst classification performance. The classification performance of DNN, LSTM, and MLP is quite consistent. Based on the test results, we determined to use GCMD as the binary classifier for the abnormal sound detection of washing machines.

As can be seen from Table 7, among all trained neural networks, convolutional neural networks have the best classification performance, and the classification accuracy of most neural networks is above 90%. This indicates that the neural network method of deep learning has a significant effect on the recognition of foreign sounds in washing machines.

5.4 Verification of robustness

The model that is trained by the data set collected and produced from the production line has a general generalization capability. The ambient noise is sometimes stronger when the actual detection is carried out on the production line, which will affect the discrimination ability of the model. To detect the generalization ability of abnormal sound recognition of home appliances in a real noisy environment for our training model, we added white noise to the sample signal before training. Adding noise to datasets can be used to improve the learning difficulty of neural networks and verify the robustness of the GCMD network [27]. It was found by experiment that the recognition capability of the GCMD network decreased only by 2.53 dB when the signal-to-noise ratio was around 10 dB, which indicates that the GCMD model still had a strong ability to generalize when added noise signals were strong.

Fig. 19 and Fig. 20 are sonograms obtained by adding a 10 dB signal-to-noise signal to a single normal sound sample and an abnormal sound sample. The blue waveform in the sonogram represents the time domain signal of the original data, and the orange waveform represents the time domain signal with noise.

As it can be seen that the noise signal has preserved the basic characteristics of the original data, but the local features changed greatly, which result in the training difficulty of the network. This is done to allow the training set data to be closer to the sound data of a real scene, thereby training the network model with a better generalization ability.

We use the original data set, the data set with a signal-to-noise ratio of 10, and the augmented training set to train the GCMD. Table 8 shows the training results of the GCMD model for these several datasets. It can be seen that the test accuracy is 0.872 of GCMD trained by the original dataset with noise. It is shown that the GCMD model has good noise robustness and overall classification performance. The model trained by an augmented training set has a higher generalization ability, which is close to 1. This shows that an augmented dataset of the abnormal sound of washing machines helps improve GCMD's performance.

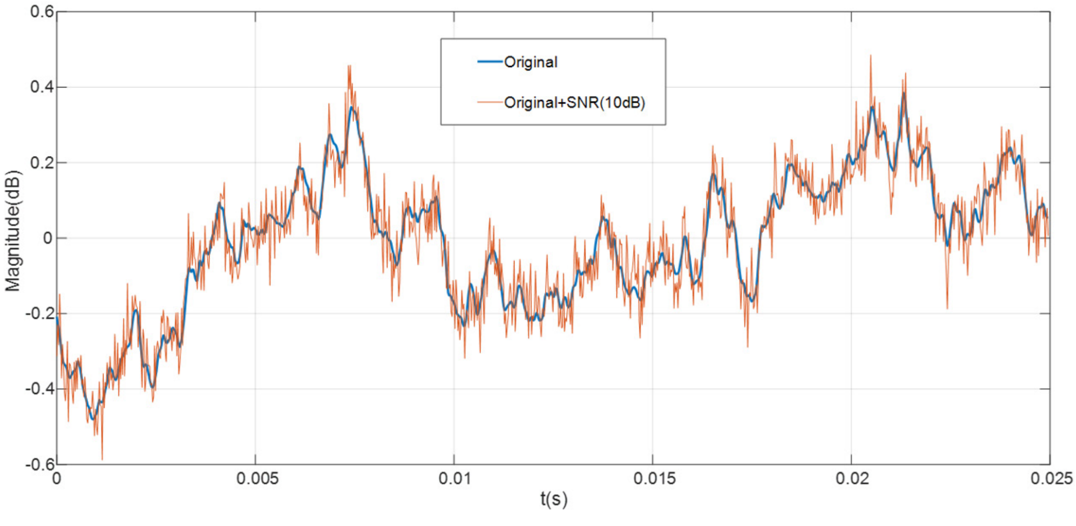


Fig. 19 Comparison of waveforms of normal data with and without noise

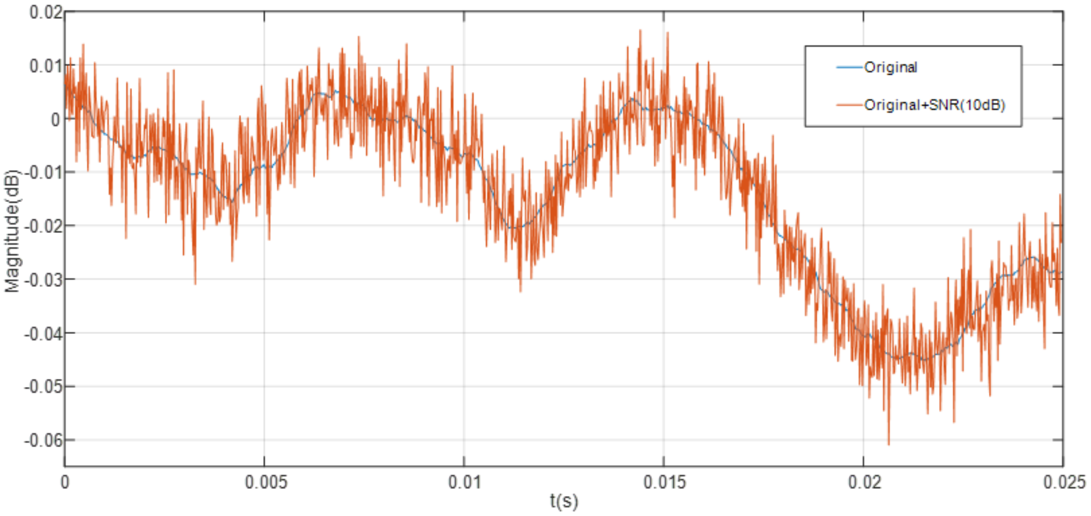


Fig. 20 Comparison of waveforms with and without noise in abnormal data

Table 8 Cross-validation accuracy for each dataset of GCMD

Dataset	Accuracy
Original dataset	0.703
Original dataset with SNR	0.872
Augmented assessment set	0.989

6 Conclusions

In this paper, aiming at the phenomenon of low model accuracy and poor generalization ability in the mechanical fault detection technology based on the deep learning method in the industrial field, we find that the mechanical fault audio data shortage problem commonly exists in the neural network. In this paper, the problem of how to improve the performance of neural networks in mechanical audio fault diagnosis is studied. In this paper, based on data augmentation technology, a model of mechanical fault sound diagnosis and recognition is proposed, which is a global average convolutional neural network GCMD applied to mechanical fault sound recognition. According to the remarkable feature extraction ability and translation invariance of CNN network architecture, we can learn the abnormal sound features of mechanical equipment in case of failure, to realize the purpose of equipment failure diagnosis. To solve the problem of data scarcity, the SBSDA data augmentation method is proposed. Based on the spectrum characteristics of log Mel and the theory of spectrum data augmentation technology, the original data is preliminarily augmented by frame selection and repeated sampling of the audio spectrum with constant or transform window size. The original log Mel spectrum of mechanical fault sound is applied to CapsuleGAN, and the fault audio data with the diversity of original data sample types are synthesized by generating the supervised learning mode of the confrontation generation of the confrontation network. A kind of artificial neural network model MFS-CapsuleGAN based on capsule generation countermeasure network (CapsuleGAN) is proposed to solve the imbalance of the data set and further improve the generalization ability of GCMD. Three kinds of augmented datasets of open mechanical fault audio data sets are generated in the experiment, and the accuracy is evaluated on the augmented data set by training GCMD neural network. Compared with the original data set, the accuracy of the model is improved by 12.3% ~ 23.7%, and the performance improvement effect is significant, which proves the feasibility and effectiveness of the MFS-CapsuleGAN data augmentation effect. In addition, the generalization ability of the GCMD network is tested by adding the data set of the background noise signal. The fluctuation range is within 0.117, which shows that the GCMD network has good robustness. All the experiments show that the augmented data set can improve the generalization ability of the model compared with the original data set. Through this method, the deep learning method is used to realize the different sound recognition of mechanical products, and it is possible to establish a perfect automatic detection production line.

Acknowledgement

This research is based upon work supported by Beijing shape data is collected Natural Science Foundation and Fengtai Rail Transit Frontier Research Joint Fund L191009, the National Natural Science Foundation of China (No. 61877002), and the scientific research program of Beijing Municipal Education Commission KZ202110011017.

References

[1] Xiao, J., Zhou, Z. "Research progress of RNN language model", In: 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, 2020, pp. 1285–1288. ISBN 978-1-7281-7006-0 https://doi.org/10.1109/ICAICA50127.2020.9182390

[2] Zhang, S. "Language processing model construction and simulation based on Hybrid CNN and LSTM", Computational Intelligence and Neuroscience, 2021, 2578422, 2021. https://doi.org/10.1155/2021/2578422

[3] Yao, H., Wang, Z., Wu, Y., Zhang, Y., Miao, K., Cui, M., Ao, T., Zhang, J., Ban, D., Zheng, H. "Intelligent sound monitoring and identification system combining triboelectric nanogenerator-based self-powered sensor with deep learning technique", Advanced Functional Materials, 32(15), 2112155, 2022. https://doi.org/10.1002/adfm.202112155

[4] Srivastava, S., Wu, H.-H., Rulff, J., Fuentes, M., Cartwright, M., Silva, C., Arora, A., Bello, J. P. "A study on robustness to perturbations for representations of environmental sound", In: 2022 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, 2022, pp. 125–129. ISBN 978-1-6654-6799-5 https://doi.org/10.23919/EUSIPCO55093.2022.9909557

[5] Liu, S., Chen, J., He, S., Shi, Z., Zhou, Z. "Subspace Network with Shared Representation learning for intelligent fault diagnosis of machine under speed transient conditions with few samples", ISA Transactions, 128, pp. 531–544, 2022. https://doi.org/10.1016/j.isatra.2021.10.025

[6] Nolasco, I., Singh, S., Vidaña-Villa, E., Grout, E., Morford, J., Emmerson, M. G., Jensens, F. H., Kiskin, I., Whitehead, H., Strandburg-Peshkin, A., Gill, L., Pamuła, H., Lostanlen, V., Morfi, V., Stowell, D. "Few-shot bioacoustic event detection at the DCASE 2022 challenge", In: Lagrange, M., Mesaros, A., Pellegrini, T., Richard, G., Serizel, R., Stowell, D. (eds.) Proceedings of the 7th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2022), Nancy, France, 2022, pp. 136–140. ISBN 978-952-03-2677-7

[7] Jaiswal, A., AbdAlmageed, W., Wu, Y., Natarajan, P. "CapsuleGAN: Generative adversarial capsule network", [preprint] arXiv, arXiv:1802.06167v7, 02 October 2018. https://doi.org/10.48550/arXiv.1802.06167

[8] Jiang, Y., Li, C., Li, N., Feng, T., Liu, M. "HAASD: A dataset of household appliances abnormal sound detection", In: Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence, Shenzhen, China, 2018, pp. 6–10. ISBN 9781450366069 https://doi.org/10.1145/3297156.3297186

[9] Koizumi, Y., Saito, S., Uematsu, H., Harada, N., Imoto, K. "ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection", In: 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 2019, pp. 313–317. ISBN 978-1-7281-1124-7 https://doi.org/10.1109/WASPAA.2019.8937164

[10] Purohit, H., Tanabe, R., Ichige, K., Endo, T., Nikaido, Y., Suefusa, K., Kawaguchi, Y. "MIMII Dataset: Sound dataset for malfunctioning industrial machine investigation and inspection", In: Mandel, M., Salamon, J., Ellis, D. P. W. (eds.) Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), New York, NY, USA, 2019, pp. 209–213. ISBN 978-0-578-59596-2 https://doi.org/10.33682/m76f-d618

[11] Zheng, T., Song, L., Wang, J., Teng, W., Xu, X., Ma, C. "Data synthesis using dual discriminator conditional generative adversarial networks for imbalanced fault diagnosis of rolling bearings", Measurement, 158, 107741, 2020. https://doi.org/10.1016/j.measurement.2020.107741

[12] Huynh, P.-H., Nguyen, V. H., Do, T.-N. "Enhancing gene expression classification of support vector machines with generative adversarial networks", Journal of Information and Communication Convergence Engineering, 17(1), pp. 14–20, 2019. https://doi.org/10.6109/jicce.2019.17.1.14

[13] Back, M.-K., Yoon, S.-W., Lee, S.-B., Lee, K.-C. "Improving Fidelity of Synthesized Voices Generated by Using GANs", KIPS Transactions on Software and Data Engineering, 10(1), pp. 9–18, 2021. https://doi.org/10.3745/KTSDE.2021.10.1.9

[14] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. "Generative adversarial networks", Communications of the ACM, 63(11), pp. 139–144, 2020. https://doi.org/10.1145/3422622

[15] Sun, C., Zhang, X., Meng, H., Cao, X., Zhang, J. "AC-WGAN-GP: Generating Labeled Samples for Improving Hyperspectral Image Classification with Small-Samples", Remote Sensing, 14(19), 4910, 2022. https://doi.org/10.3390/rs14194910

[16] Radford, A., Metz, L., Chintala, S. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks", [preprint] arXiv, arXiv:1511.06434v2, 07 January 2016. https://doi.org/10.48550/arXiv.1511.06434

[17] Zhou, Y., Sun, J. "Speech Recognition Using Double Data Augmentation Strategy", In: 2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 2022, pp. 715–719. ISBN 978-1-6654-2208-6 https://doi.org/10.1109/ITAIC54216.2022.9836801

[18] Donahue, C., McAuley, J., Puckette, M. "Adversarial audio synthesis", [preprint] arXiv, arXiv:1802.04208v3, 09 February 2019. https://doi.org/10.48550/arXiv.1802.04208

[19] Arjovsky, M., Chintala, S., Bottou, L. "Wasserstein generative adversarial networks", Proceedings of Machine Learning Research, 70, pp. 214–223, 2017.

[20] Zhang, T., Li, Z., Zhu, Q., Zhang, D. "Improved procedures for training primal wasserstein gans", In: 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), Leicester, UK, 2019, pp. 1601–1607. ISBN 978-1-7281-4035-3 https://doi.org/10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00286

[21] Yamamoto, R., Song, E., Kim, J.-M. "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram", In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6199–6203. ISBN 978-1-5090-6632-2
https://doi.org/10.1109/ICASSP40776.2020.9053795

[22] Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., Le, Q. V. "SpecAugment: A simple data augmentation method for automatic speech recognition", [preprint], arXiv, arXiv:1904.08779v3, 03 December 2019.
https://doi.org/10.48550/arXiv.1904.08779

[23] Liu, Y., Neophytou, A., Sengupta, S., Sommerlade, E. "Cross-modal spectrum transformation network for acoustic scene classification", In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 830–834. ISBN 978-1-7281-7606-2
https://doi.org/10.1109/ICASSP39728.2021.9414779

[24] Nenov, R., Nguyen, D.-K., Balazs, P. "Faster Than Fast: Accelerating the Griffin-Lim Algorithm", In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1–5. ISBN 978-1-7281-6328-4
https://doi.org/10.1109/ICASSP49357.2023.10097224

[25] Zhang, P., Pang, X., Kibalya, G., Kumar, N., He, S., Zhao, B. "GCMD: Genetic correlation multi-domain virtual network embedding algorithm", IEEE Access, 9, pp. 67167–67175, 2021.
https://doi.org/10.1109/ACCESS.2021.3076916

[26] Gomez, A. N., Zhang, I., Kamalakara, S. R., Madaan, D., Swersky, K., Gal, Y., Hinton, G. E. "Learning sparse networks using targeted dropout", [preprint] arXiv, arXiv:1905.13678v5, 09 September 2019.
https://doi.org/10.48550/arXiv.1905.13678

[27] Yu, S., Chen, M., Zhang, E., Wu, J., Yu, H., Yang, Z., Ma, L., Gu, X., Lu, W. "Robustness study of noisy annotation in deep learning based medical image segmentation", Physics in Medicine & Biology, 65(17), 175007, 2020.
https://doi.org/10.1088/1361-6560/ab99e5