IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 19, 2022, accepted June 1, 2022, date of publication June 8, 2022, date of current version June 13, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3180725

Multimodal Fusion Convolutional Neural Network With Cross-Attention Mechanism for Internal Defect Detection of Magnetic Tile

HOUHONG LU, YANGYANG ZHU, MING YIN, GUOFU YIN, AND LUOFENG XIE
School of Mechanical Engineering, Sichuan University, Chengdu 610065, China

Corresponding author: Luofeng Xie (xielf@scu.edu.cn)

This work was supported in part by the Postdoctoral Science Foundation of China under Grant 2021M692322, in part by the Department of Science and Technology of Sichuan Province under Grant 2020ZDZX0003, and in part by the Fundamental Research Funds for the Central Universities under Grant 2021SCU12146.

ABSTRACT The internal defect detection of magnetic tile is extremely significant before mounting. Currently, this task is completely realized by manual operation in the magnetic tile manufacturing industry, which results in inefficiency and diseconomy. In this work, we develop an intelligent system based on the acoustic sound for internal defect detection of magnetic tile to overcome these drawbacks. Due to the non-Gaussian and non-stationary characteristics of the acoustic sound, adopting the single modality of the data for internal defect detection of magnetic tile cannot achieve good accuracy. Therefore, we design a multimodal fusion convolutional neural network (MMFCNN) for internal defect detection of magnetic tile. We train the network in an end-to-end way. Our proposed MMFCNN consists of three blocks, i.e., feature extraction block, feature fusion block and internal defect detection block, whose purposes are to extract features from generated modal data, fuse multimodal feature maps and analyze whether the magnetic tile has internal defects, respectively. Moreover, to realize the information interaction and emphasize more representative information at feature extraction stage, we propose a novel attention mechanism, i.e., cross-attention mechanism. Extensive experimental results demonstrate our proposed MMFCNN is effective for internal defect detection of magnetic tile. Our code is available at https://github.com/Clarkxielf/Multimodal-Fusion-Convolutional- Neural-Network-for-Internal-Defect-Detection-of-Magnetic-Tile.

INDEX TERMS Magnetic tile, internal defect detection, convolutional neural network (CNN), feature fusion, cross-attention mechanism.

I. INTRODUCTION

Magnetic tile is a kind of arc permanent magnet, which is the core component of the permanent magnet motor [1]. With the rapid development of automation technology, abundant component magnet motors are widely used in automation equipment and intelligent device. Therefore, its quality plays a decisive role in the performance and service life of electromechanical products. As a kind of clean and cheap energy, magnetic tile has not only a wide variety but also growing global market demand, especially in the field of electric vehicles. The defects of magnetic tile are mainly divided into two categories: external defects and internal defects. As so far, researchers have proposed many methods to detect external defects based on machine vision technologies [2], [3]. On the contrary, internal defects are invisible, bringing new challenges to their detection. There are many factors causing the internal defects of magnetic tiles, such as uneven raw materials, thermal shock and rapid cooling in the production process. Currently, internal defects of magnetic tile are identified by experienced workers through listening intently to the excited sound in the magnetic tile manufacturing industry. But such process is extremely risky. For the magnetic tile manufacturers, once the internal defects of the sold magnetic tile are detected by the user, this batch of magnetic tiles will be scrapped and recycled, which would cause serious economic losses. More seriously, if the magnetic tile with internal defects is used, it is likely to cause safety accidents and casualties. Therefore, developing an automation system to detect internal

The associate editor coordinating the review of this manuscript and approving it for publication was Guillermo Valencia-Palomo.

defects is becoming increasingly urgent in the magnetic tile industry.

Nowadays, with the development of science and technology, abundant non-destructive testing technologies have been successfully developed for internal defects by scientists around the world, such as ultrasound, infrared imaging, acoustic emission and X-ray diffraction tomography [4]. Although these technologies have attained great success and are widely applied in many non-destructive testing scenarios, they are too costly to operate in an automatic way to match the agile manufacturing process for different kinds of magnetic tiles. Inspired by the manual operation in the magnetic tile manufacturing industry, we utilize the acoustic sound for internal defect detection of magnetic tile. This is because the characteristics of the acoustic sound for an object are closely linked to its physical structure vibration.

However, the acquired acoustic sound is generally nonlinear, non-Gaussian and non-stationary, which can seriously hinder the extraction and identification of the signal features regarding internal defects, whereas these meaningful features are usually too weak to be discovered [5]. Therefore, many algorithms are proposed to process acquired acoustic sounds, such as wavelet packet analysis (WPT) [6], hidden Markov model (HDM) [7], principal component analysis (PCA) [8] and variational mode decomposition (VMD) [9]. But those algorithms need to design hand-crafted features, which requires complex mathematical operations and a certain understanding of the extracted signals as well as a wealth of signal processing knowledge. More importantly, those specially designed hand-crafted features generally work well for specific signal and fault scenarios and are probably not applicable for diverse types of time-series and different operating conditions. To address this issue, it is superior to design an end-to-end algorithm to analyze acoustic sounds of objects without much expert knowledge. Therefore, deep learning (DL) [10]–[13] is always a good choice for such a situation.

As a special machine learning model, deep learning techniques are structured by a stack of multiple layers of nonlinear processing units. It shows excellent performance in many fields, e.g., image classification [14], target recognition [15], semantic segmentation [16], natural language processing [17], machine translation [18], and so on. Compared with the traditional machine learning algorithms, DL techniques are capable of intelligently learning underlying features from large and diverse data, which escapes from the dilemma of hand-crafted feature design. Especially, convolutional neural networks (CNNs) are the most widely used to extract meaningful features. From AlexNet [19] to ResNet [20], the depth of CNNs becomes deeper and deeper, and the number of parameters becomes larger and larger. AlexNet uses Rectified Linear Unit (ReLU) [21] to replace the traditional activation function to solve the gradient dispersion problem, and adopts Dropout to prevent overfitting of the model. VGG [22] stacks multiple small convolution kernels to replace a large convolution kernel, which can significantly improve the learning ability of the network. This is because the nonlinear ability of multiple small convolutional kernels is stronger than that of a larger convolutional kernel. GoogleNet [23] performs multiple convolutional operations with different kernel sizes on features in parallel to learn multi-scale representation information. ResNet introduces the residual shortcut to solve the gradient disappearance problem of deep network, which strengthens the information interaction between adjacent residual blocks. Later, more and more lightweight CNNs [24], [25] are proposed to reduce the inference time of the model without compromising the performance. Although the aforementioned CNNs show good performance in classification tasks, it is not applicable to the internal defect detection of magnetic tile based on the acoustic data because of the non-Gaussian and non-stationary characteristics of the acoustic sound. Moreover, the unknown size, shape and location of internal defects also increase the difficulty in extracting effective features embedded in acoustic sound. Therefore, only extracting features from time-domain acoustic sound cannot completely characterize internal defects of the magnetic tile since the acoustic sound in time domain only reflects the fluctuation of sound energy over a period of time.

Currently, researchers have done a lot of studies on the classification of multimodal fusion based on deep learning for the time-series signal [26]–[32]. According to different inputs, multimodal fusion is divided into two categories. The first one is that inputs include various signals, i.e., voice, text, image or data from different sensors. For example, Wang *et al.* [32] proposed a new deep learning-based prognostics framework for predicting the remaining useful life of machinery, which utilizes monitoring data from different sensors as the inputs of the prognostics network so as to integrate the complete degradation information. The other is that inputs are the multiple transformed signals of one signal. Ahmad *et al.* [29] proposed two efficient multimodal fusion networks for electrocardiogram (ECG) heart beat classification, whose inputs are images of Gramian Angular Field, Recurrence Plot and Markov Transition Field. Liang *et al.* [31] proposed a new methodology of parallel convolutional neural network (P-CNN) for bearing fault identification, which is capable of extracting features from time domain and time-frequency domain of the raw vibration signal. However, mostly previous works only simply stack extracted features of each modality for fusion, without considering the degree of importance among features. Although few researchers assign weights to the features of each modality, they ignore the differences of cross-modal features.

Therefore, based on the latter fusion method, a novel CNN framework termed MMFCNN is proposed for internal defect detection of magnetic tile in this article. Its inputs are signals of the raw time domain, frequency domain gained by fast Fourier transform (FFT) and time-frequency domain yielded by spectrogram transform. In the proposed MMFCNN,

a multi-branch feature extraction strategy is developed to learn high-dimensional representations from time domain, frequency domain and time-frequency domain of the acoustic data. Then, the cross-attention mechanism is proposed into MMFCNN to realize the information interaction among each branch and emphasize more representative features at feature extraction stage. Next, feature fusion block integrates the high-level information from feature extraction block. Finally, the integrated representations are fed into an internal defect detection block for internal defect detection of magnetic tile. The main contributions of this article can be summarized as follows.

1) MMFCNN architecture is proposed by three parallel CNN branches, which can extract efficient representations from three different domains of the raw acoustic data.
2) Cross-attention mechanism is proposed into the MMFCNN to focus on more important features and interact information among branches.
3) An intelligent system for internal defect detection of magnetic tile is developed. This system can automatically and efficiently classify magnetic tiles. The detection speed of the system shall reach 40 magnetic tiles per minute at least. Therefore, it has great practical value for the magnetic tile industry.

The rest of the article is summarized as follows. In Section II, details of the proposed MMFCNN are elaborated. In Section III, the method of internal defect detection of magnetic tile is given. Section IV presents the experimental results, and Section V concludes this article.

II. PROPOSED MMFCNN

In this work, our goal is to compose an intelligent network, which is capable of disclosing the mapping relationship between acoustic data and defect labels (whether there are internal defects in magnetic tiles). However, due to the acoustic sound being non-Gaussian and non-stationary, adopting the single modality of the acoustic sound for internal defect detection of magnetic tile cannot achieve good accuracy. To overcome this problem, we design a multi-branch neural network to extract features from time domain, frequency domain and time-frequency domain.

The architecture of the proposed MMFCNN is illustrated in Fig. 1, which consists of feature extraction block, feature fusion block and internal defect detection block. The raw acoustic data collected by a sound acquisition sensor, are transformed by Fourier transform and spectrogram. The raw acoustic data together with two kinds of transformed data are first input into the feature extraction block to learn multidimensional representations. Meanwhile, the high-dimensional representations are fed into feature fusion block to fuse the differently useful information of multimodal data. Finally, we input the fused features into the internal defect detection block to analyze whether the magnetic tile has internal defects. The details of MMFCNN are described as follows.

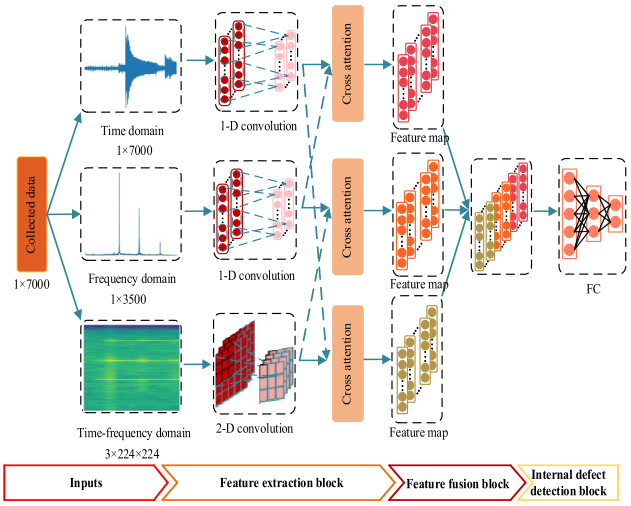

FIGURE 1. The architecture of MMFCNN.

A. FEATURE EXTRACTION BLOCK

The feature extraction block is structured by three streams and each stream consists of several CNN layers. And the architecture of each stream in MMFCNN is shown in Table 1. In particular, to emphatically concern the important information and effectively fuse the complementary features, the cross-attention mechanism is established behind the convolutional module.

TABLE 1. The architecture of each stream in MMFCNN.

Name	Network1	Network2
Input layer	1×7000 &1×3500	3×224×224
Conv1_X	Filter: 3×1, 64, stride:1 BN + ReLU 3×3 max pooling, stride:2	Filter: 11×11, 64, stride:4 BN + ReLU 3×3 max pooling, stride:2
Conv2_X	Filter: 3×1, 192, stride:1 BN + ReLU 3×3 max pooling, stride:2	Filter: 5×5, 192, stride:2 BN + ReLU 3×3 max pooling, stride:2
Conv3_X	Filter:3×1, 384, stride:1 BN + ReLU	Filter: 3×3, 384, stride:1 BN + ReLU
Conv4_X	Filter:3×1, 256, stride:1 BN + ReLU	Filter: 3×3, 256, stride:1 BN + ReLU
Conv5_X	Filter:3×1, 256, stride:1 BN + ReLU 3×3 max pooling, stride:2	Filter: 3×3, 256, stride:1 BN + ReLU 3×3 max pooling, stride:2

1) CONVOLUTIONAL MODULE

In the constructed MMFCNN, the architecture of the convolutional module is established by a series of CNN layers. The convolutional layer firstly utilizes several learnable convolutional kernels to convolve the input data, and then, applying an elementwise nonlinear activation function on the outputs of convolution operations. To avoid the overfitting of this model, batch normalization (BN) [33] is implanted in the convolutional layer. Through those three operations, different feature maps can be obtained in a convolutional

layer. Mathematically, it can be expressed as follows.

$$x^l = \sigma_r(\phi(\omega(x^{l-1}))) \tag{1}$$

where x^l represents the feature map of the *lth* convolutional layer, $\sigma_r(\cdot)$ is the rectified linear unit (ReLU) activation function, $\phi(\cdot)$ is the batch normalization operation, and $\omega(\cdot)$ denotes the convolutional operation.

Due to the signals in the time domain and frequency domain being one-dimensional (1-D) sequences, so 1-D convolution is used to extract features. On the other hand, after the signal is transformed by spectrogram, the output is an RGB image containing time domain and frequency domain information. Thus, 2-D convolution is utilized to learn the representation of the spectrogram. The principle of one-dimensional convolution kernel and two-dimensional convolution is the same. For convenience, we denote the time domain, the frequency domain and the time-frequency domain, respectively, as D_T, D_F and D_{T-F}. Let $x_T^{l-1} \in \mathbb{R}^{L\times1\times C}$ ($x_F^{l-1} \in \mathbb{R}^{L\times1\times C}$, $x_{T-F}^{l-1} \in \mathbb{R}^{W\times H\times C}$) and $k_1^l \in \mathbb{R}^{K\times1\times C\times N}$ ($k_2^l \in \mathbb{R}^{K\times K\times C\times N}$) represent the input volume in D_T (D_F, D_{T-F}) and 1-D (2-D) convolutional kernel, where L is the length of the input volume, W and H denote the width and height, N represents the number of the convolutional kernel, $K\times1$ ($K\times K$) is the convolutional kernel size. The output of convolution of the *lth* convolutional layer can be calculated by

$$u^l = k^l * x^{l-1} + b = \sum_{c=1}^{C} k_c^l * x_c^{l-1} + b^l \tag{2}$$

where b^l denotes the bias, $*$ denotes the convolutional operation, and C represents the number of input channels.

To extract the main features of convolutional operation and increase the receptive field, the pooling layer is optionally used behind the convolutional layer. As an independent neural layer, the pooling operation has no parameters, and is used to filter out unnecessary characteristics and preserve vital representations. As a result, the obtained feature maps cover the significant information of the raw data. Mathematically, the *nth* feature map of the *lth* pooling layer y_n^l cloud be expressed by

$$y_n^l = pool(x_n^l, k, s) \tag{3}$$

where x_n^l is the *nth* output feature map of the *lth* convolutional layer, i.e., $pool(\cdot)$ is the max pooling operation, k is the pooling kernel size, and s represents the stride of the pooling kernel.

It is worth noting that the dimensions of the feature maps outputted by the convolutional module of three branches are not the same, which will bring difficulties to the subsequent information interaction and feature fusion. Therefore, we flatten the feature maps generated by the 2-D convolution module along the channel, and then do a 1-D convolution operation. The output feature map F_z can be formulated as

$$F_z = conv1d(f(x_{T-F}^l)) \tag{4}$$

where $f(\cdot)$ denotes the flattening operation.

2) CROSS-ATTENTION MECHANISM

In essence, the attention mechanism in CNN is similar to the human selective visual attention mechanism, and the core goal is to select the information that is more critical to the current task from numerous information. Specifically, introducing an attention mechanism in CNN is to emphasize more representative features that are relevant to the internal structure of magnetic tile while restraining inessential information. On the other hand, to realize the information interaction among feature extraction branches, a novel module named the cross-attention mechanism is introduced into our designed MMFCNN. Moreover, due to the differences among cross-modal features, the cross-attention mechanism can make incompatible features align in fused feature space. As shown in Fig. 2, it consists of two blocks: channel-wise attention and feature interaction mechanism [34].

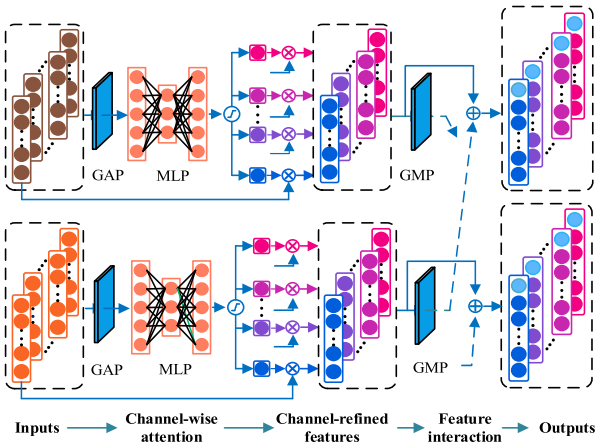

FIGURE 2. Illustration for cross-attention mechanism in MMFCNN.

a: CHANNEL-WISE ATTENTION

In CNN, each channel of the feature maps is the activation response corresponding to the convolution kernel, and introducing channel-wise attention mechanism into CNN can be regarded as the process of selecting semantics [35], which learns the weight of each channel and improves the representation performance of convolution features by suppressing irrelevant features. In channel-wise attention, firstly, an adaptive average pooling is carried out behind the convolutional module. Then, it is forward into a multilayer perception (MLP) with two layers, which yields a feature vector. Last, the output feature vector is fed into the sigmoid activation function to obtain the channel-wise attention vector. It can be calculated by

$$V_{ca} = \sigma_s(MLP(pool_a(F_x))) \tag{5}$$

where V_{ca} is the channel-wise attention vector, $\sigma_s(\cdot)$ represents the sigmoid activation function, $pool(\cdot)$ denotes the global average-pooling (GAP), and $F_x \in \mathbb{R}^{B\times I\times C}$ is the output feature map of the convolutional module.

b: FEATURE INTERACTION MECHANISM

The global feature of the single modality is crucial for classification. Introducing feature interaction mechanism is to take advantage of this global information, so that each branch contains important information of other branches. Given feature maps $F_x, F_y, F_z \in \mathbb{R}^{B\times I\times C}$ extracted from three kinds of modal data, where B is the batch-size, C is the number of channels and I is the length of each feature map. We do feature interaction for any two of the three kinds of features. Due to the operation of feature interaction is asymmetric, selecting features of three domains to do feature interaction, which eventually leads to the fusion features containing redundant information. For convenience, just take F_x and F_y as examples. Firstly, F_y is aggregated by channel-wise max-pooling to gain the globally significant feature, then the feature is concatenated with F_x at the same level. It can be expressed by

$$\overline{F_x} = \psi(F_x, pool_m(F_y)) \tag{6}$$

where $\overline{F_x} \in \mathbb{R}^{B\times(I+1)\times C}$, $\psi(\cdot)$ represents the concatenation operation, $pool_m(\cdot)$ is the global max-pooling (GMP).

B. FEATURE FUSION BLOCK

As mentioned above, a single acoustic sound cannot well realize the task of internal defect detection of magnetic tile. Therefore, a feature fusion strategy that can make use of the differently useful information of the generated modal data, is embedded into our proposed MMFCNN. For multi-dimension feature maps, there are many ways for multimodal feature fusion, including max, mean, sum and concatenation operation.

For max fusion operator, it calculates the max values of three modal feature maps at the same spatial locations a of the cth channel. The max fusion operator $\Gamma^{a,c}_{\max}$ can be expressed as

$$\Gamma^{a,c}_{\max} = \max\{\overline{F_x}^{a,c}, \overline{F_y}^{a,c}, \overline{F_z}^{a,c}\} \tag{7}$$

For mean fusion operator, it calculates the mean values of three modal feature maps at the same spatial locations a of the cth channel. The mean fusion operator $\Gamma^{c,a}_{mean}$ is expressed as

$$\Gamma^{a,c}_{mean} = \frac{1}{3}\sum(\overline{F_x}^{a,c}, \overline{F_y}^{a,c}, \overline{F_z}^{a,c}) \tag{8}$$

For concatenation fusion operator, it means the high-dimensional feature maps of multimodal data are stacked along the channel direction. Mathematically, the concatenation fusion operator can be expressed as

$$\begin{aligned}\Gamma_{cat} &= \psi(\overline{F_x}, \overline{F_y}, \overline{F_z}) \\ &= \{\overline{F_x}^1, \ldots, \overline{F_x}^C, \overline{F_y}^1, \ldots, \overline{F_y}^C, \overline{F_z}^1, \ldots, \overline{F_z}^C\}\end{aligned} \tag{9}$$

where Γ_{cat} is the output of the concatenation fusion operator of three branches feature maps, $\psi(\cdot)$ is the concatenate operation, $\overline{F_x}$, $\overline{F_y}$ and $\overline{F_z}$ represent the output feature maps of cross-attention mechanism, and C is the number of channels.

For sum fusion operator, it calculates the sum values of feature maps at the same spatial locations. Since the importance of each modal feature map is unclear, we assign learnable parameters α, β, γ to feature maps of three generated modal data, which represents the weight of each feature. Mathematically, it can be expressed as

$$\Gamma^{a,c}_{sum} = \sum_{c=1}^{C}(\alpha\overline{F_x}^{a,c}, \beta\overline{F_y}^{a,c}, \gamma\overline{F_z}^{a,c}) \tag{10}$$

where a represents the spatial location and c is the cth channel of feature maps. The detail of learnable parameters α, β, γ is described in Section IV (D).

C. INTERNAL DEFECT DETECTION BLOCK

The specifically designed internal defect detection block consists of four fully connected layers (FCLs). And these four FCLs contains 2048, 512, 128 and 2 neurons, respectively. The first three fully connected layers are associated with the Dropout and ReLUs. The feature map of feature fusion block is then flattened to be fed to four FCLs.

D. LOSS FUNCTION

Essentially, the internal defect detection of magnetic tile is a binary classification problem. Therefore, the binary cross-entropy loss is chosen as the loss function. It is defined as

$$\ell = -(p\log\bar{p} + (1-p)\log(1-\bar{p})) \tag{11}$$

where $\bar{p}$ denotes the probability that the predicted result is a positive example (without internal defects), and p represents the label of the sample. If the sample is a positive example, the value is 1; otherwise, the value is 0.

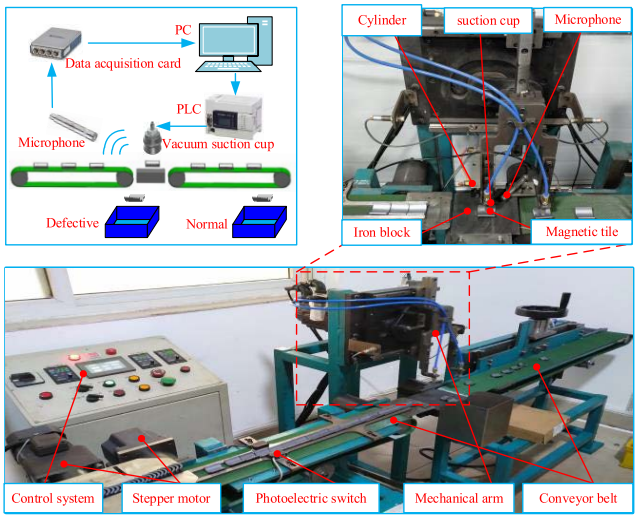

FIGURE 3. Scheme of internal defect detection system.

III. METHODS

A. SYSTEM SETUP

As shown in Fig. 3, we designed an intelligent detection system for internal defects of magnetic tiles, which can

automatically collect sound and send it to the computer for prediction, and then feedback the prediction results to the classification system to classify the magnetic tile with or without internal defects. This system consists of five parts, namely, transportation system, excitation system, sound acquisition system, internal defect detection system, and sorting device. The composition of each part is as follows: the transportation system consists of three parallel conveyor belts. The first two conveyor belts carry the magnetic tiles in an upright and transverse posture respectively, and the third conveyor belt transports the sampled magnetic tile to the designated position for sorting. The excitation system is essentially a mechanical arm, which is responsible for grasping the magnetic tile to about two centimeters height and then falling to collide with the iron block to generate sound. The sound acquisition system is a data acquisition card with a microphone. The internal defect detection system is an application software, and its detection process is to call the prediction program based on MMFCNN. Last, the sorting device is composed of two cylinders, which remove broken magnetic tiles (the magnetic tile with obvious internal crack is easy to be broken after colliding) and magnetic tiles with internal defects, respectively.

The working principle of this system is summarized as follows. At first, the mechanical arm goes down to grab the transverse magnetic tiles to a fixed height. Then, sound acquisition system collects the sound generated by the magnetic tile colliding with the iron block after falling. Finally, the collected sound is input to the designed model for prediction, and the prediction results are fed back to the sorting device for classification.

B. DATASET CONSTUCTION

Before training a deep neural network, it is essential to obtain data and label the corresponding labels. However, the internal defects of magnetic tile are not as obvious as the surface defects. In industry, the internal defect detection process in magnetic tile mainly depends on the hearing of experienced workers. They distinguish magnetic tiles with internal defects from the sounds generated by the magnetic tile colliding with the iron block.

To realize the trained model can be well applied to the designed equipment, we use the designed device to sample the acoustic data of the magnetic tiles, which are labelled by experienced workers in advance. As for sampling parameters, the sampling frequency is set to be 40 kHz and 7000 data points are recorded for each sound. In the end, we obtained 1241 magnetic tile samples, including 730 samples with internal defects and 511 normal samples. The split of the dataset is shown in Table 2. Furthermore, the sample with the internal defect was labelled as ''Defective,'' on the contrary, it was labelled as ''Normal.''

C. DATA PROCESSING

Data processing is critical to the model training. In this work, there are three main data processing methods, i.e., data normalization, FFT and spectrogram transform. Data normalization is helpful to adjust the learning rate and accelerate the convergence speed. And the data transformed by FFT and spectrogram will be used as the input of the proposed MMFCNN together with the raw acoustic data. The details of these data processing methods are as follows.

TABLE 2. Distribution of experimental samples.

Dataset	Defective sample	Normal sample
Training set	583	408
Validation set	147	103

In this work, we adopt the min-max normalization method. It transforms the range of the original signal to [0,1] without changing the shape of the original signal. Mathematically, it can be expressed as follows.

$$x^* = \frac{x[n] - x[n]_{\min}}{x[n]_{\max} - x[n]_{\min}} \tag{12}$$

where $x[n]$ is the raw acoustic data, $x[n]_{\min}$ and $x[n]_{\max}$ represent the minimum and maximum values in $x[n]$.

FFT is a simplified form of discrete Fourier transform (DFT). DFT is defined as follows.

$$X(k) = DFT(x[n]) = \sum_{n=0}^{N-1} x[n] \cdot e^{-j\frac{2\pi}{N}nk} \tag{13}$$

where $X(k)$ is the DFT of kth point in $x[n]$, N denotes the length of $x[n]$. Let $\overline{N} = 2^m$ be the largest integer tending to N, and $W_{\overline{N}}^{k,n} = e^{-2\pi jnk/N}$. Duo to $W_{\overline{N}}^{k,n}$ is periodic and symmetrical, the amount of DFT calculation can be significantly reduced. This case is defined as FFT, and the FFT of kthpoint in $x[n]$ can be calculated by

$$\overline{X}(k) = FFT(x[n]) = \sum_{n=0}^{N-1} x[n] \cdot W_{\overline{N}}^{k,n} \tag{14}$$

Spectrogram transform mainly includes three steps, i.e., framing, windowing function and FFT. Framing is to divide the sound into small segments with fixed length (For example, 25 milliseconds). The signal of each frame is usually multiplied by a smooth window function to make both ends of the frame decay smoothly to zero for FFT. Here, we use Hamming window function. It can be expressed by

$$\tau(t) = \begin{cases} 0.54 - 0.46\cos\ [2\pi t/(M-1)], & 0 \le t \le M \\ 0, & other \end{cases} \tag{15}$$

where M is the length of the frame. To show the importance of each frame of data, we calculate the energy of each frame with the following formula.

$$E = \frac{|FFT(x[n]_i)|^2}{M} \tag{16}$$

where $x[n]_i$ represents the ith frame of the acoustic sound.

IV. EXPERIMENTAL RESULTS

The proposed MMFCNN is trained on 4 NVIDIA GeForce RTX 2080ti GPUs using PyTorch, a deep learning framework. The initial learning rate is 0.01 and decays by a factor of 10 every 20 epochs at the last 40 epochs. The synchronous SGD optimizer is adopted with weight 1e-5, momentum 0.9. The total epochs are 200 and the size of mini-batch is 32.

As a binary classification problem, the mapping relationship between acoustic data and internal defects of magnetic tile is relatively simple. The number of collected samples is sufficient to achieve good generalization performance, so no data augmentation technique is used. To make experimental results be more persuasive, each network is run five time. Then, the final results are presented through the mean and standard deviation. In comparative experiments (B, C, D), we don't add any attention mechanism.

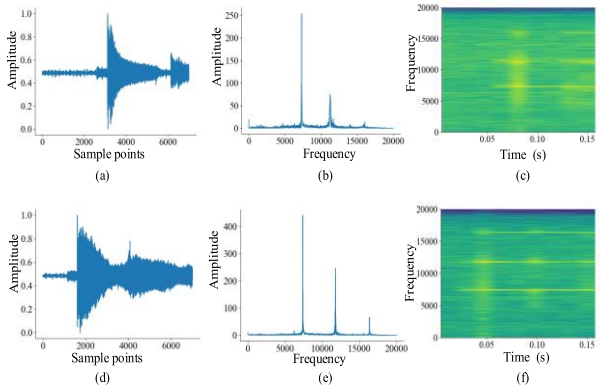

FIGURE 4. Visualization of data difference. (a)∼(c) The time-domain signal, frequency-domain signal and time-frequency-domain signal of "Defective" magnetic tile. (d)∼(f) that of "Normal" magnetic tile.

A. DATA DIFFERENCE

As shown in Fig. 4, it shows the visualization of the acoustic data of the normal and defective magnetic tiles in three kinds of domains, respectively. In the time domain, the signal represents the fluctuation of sound energy over a period of time. As can be seen from the first column in Fig. 4, it is quite difficult to distinguish the difference of acoustic data between the normal and defective magnetic tile. This situation is not conducive to achieving accurate classification. Then, we convert the time domain signal to the frequency domain and time-frequency domain. Because the signal after FFT is symmetrical, we only take advantage of half of the data to avoid information redundancy. In the frequency domain space, the signal shows the distribution of frequency of each component wave. For defective and normal magnetic tiles, the dominant frequencies of their sound signals are mainly distributed around 7500Hz, 12000Hz and 16500Hz. However, for defective ones, the curve of the frequency domain signal contains more small peaks than that of normal ones. These small peaks are caused by the magnetic tile with internal defects. The spectrogram shows the distribution relationship between energy and frequency of the acoustic sound. As can be seen from the third column in Fig. 4, there are several bright lines in the spectrogram, which represent multiple dominant frequencies of the acoustic sound and high-energy areas. By comparison, the color of the area near the bright line in the spectrum of defective magnetic tile is brighter than that of normal magnetic tile, which corresponds to the distribution of sound signal in frequency domain.

TABLE 3. Performance comparison in time domain.

Architecture	Accuracy rate (%)
AlexNet	94.08(±0.18)
VGG-16	93.36(±0.22)
ResNet-50	94.16(±0.22)
MMFCNN-A	97.12(±1.56)
MMFCNN-V	97.68(±0.18)
MMFCNN-R	97.52(±0.18)
MMFCNN	98.16(±0.22)

B. COMPARISION WITH CLASSICAL CNNS

In this article, to demonstrate the superiority of our proposed MMFCNN, we compare our model with three famous networks, i.e., AlexNet, VGG-16 and ResNet-18. These models show state-of-the-art performance in the field of image classification. Besides, three generated MMFCNNs (MMFCNN-A, MMFCNN-V and MMFCNN-R) are compared, whose backbones are aforementioned three networks. Table 3 summarizes the performance comparison results of the proposed MMFCNN and the aforementioned CNNs in the internal defect detection of magnetic tiles. As shown in Table 3, the accuracy of the proposed MMFCNN is much better than that of aforementioned CNNs, whose accuracy rate reaches 98.16%. While, the maximum accuracy rate of aforementioned CNNs is 97.68%, which demonstrates that only extracting the characteristics of sound signal in time domain cannot achieve good results in predicting the internal defects of magnetic tile and our proposed MMFCNN is relatively superior. Besides, the deeper CNNs are, the higher the accuracy rates cannot be significantly improved.

C. EFFECTIVENESS OF FEATURE FUSION

In this article, to verify that each modal data contributes to the internal defect detection of magnetic tile and feature fusion is effective, seven kinds of architectures are compared. For simplicity, the time domain, frequency domain and time-frequency domain are referred to as T, F and T-F respectively, and all combinations between them are also obtained, i.e., T+F, T+T-F, F+T-F and T+F+T-F. For the data in T, F and T-F, they are input to the single CNN for training. Moreover, these data in T+F, T+T-F and F+T-F are respectively fed into MMFCNN with two branches for training. Correspondingly, the data in T+F+T-F are input to MMFCNN with three branches for training. For the last four cases, these networks all use concatenation operation as the way of feature fusion.

TABLE 4. Accuracy rates based on different architecture.

Data mode	Architecture	Accuracy rate (%)
T	Network1	93.04(±0.61)
F	Network1	97.04(±0.22)
T-F	Network2	97.12(±0.19)
T+F	Network1+Network1	96.72(±0.33)
T+T-F	Network1+Network2	98.08(±0.19)
F+T-F	Network1+Network2	98.08(±0.33)
T+F+T-F	Network1+Network1+Network2	98.16(±0.22)

Table 4 summarizes the performance comparison results. As shown in Table 4, the highest accuracy rate is 98.16%, whose architecture uses three modal data as input. As can be observed from Table 4, the accuracy rate of feature fusion is much higher than the single network, which illustrates feature fusion is effective. However, the effect of feature fusion between time domain and frequent domain is poorer than the single network in frequency domain. This is because the features in time domain and frequency domain are inconsistent, which leads to disorder of defect information through simple feature stacking. Another obvious result is that the more modalities of data are fused, the higher the prediction accuracy is, which shows each modal data contributes to the internal defect detection. This is because each modal data describes internal defects from a different angle. Moreover, it demonstrates that feature extracted from different modalities can supplement extra information for internal defect detection of magnetic tiles.

D. FEATURE FUSION METHODS COMPARISON

In this experiment, feature fusion methods are explored. For these four fusion methods, the max operation extracts the most salient feature, the mean operation balances three types of features, the sum operation makes a suitable combination of features, and the concatenation operation integrates all defect features, which are independent of each other in the fused features. For sum fusion operation, we designed a subnetwork to regress three trainable weight parameters, i.e., α, β, γ, which were assigned to the feature maps of corresponding modes. The architecture of this subnetwork consists of four layers of processing units, i.e., a GMP and three convolutional layers. The GMP samples down the size of the feature map to 1. The subsequent three convolution layers with 1×1 kernel size, contain 256, 64 and 3 channels, respectively. Finally, three weight values are obtained through the softmax activation function.

TABLE 5. Accuracy rates of MMFCNN based on different fusion methods.

Fusion method	Accuracy rate (%)
max	98.16(±0.22)
mean	98.08(±0.33)
sum	97.60(±0.49)
concatenation	98.16(±0.22)

The experiment results are summarized in Table 5. As can be seen from Table 5, using max or concatenation operation for feature fusion achieves the best result, whose accuracy rate reaches 98.16%. The mean operation is a little better than the sum operation, the accuracy rates of which are 98.08% and 97.60%, respectively. The accuracy of these fusion methods is very close, which cannot explain which fusion method has a better effect. To illustrate the impact and generalization performance of these four fusion methods on MMFCNN, the training loss and the validation accuracy rates based on the aforementioned fusion methods are shown in Fig. 5(a) and Fig. 5(b). As shown in Fig. 5, our proposed algorithm based on these four fusion methods all converges after 200 epochs. Especially, the concatenation operation converges faster than others, and obtains higher accuracy on the verification set after convergence.

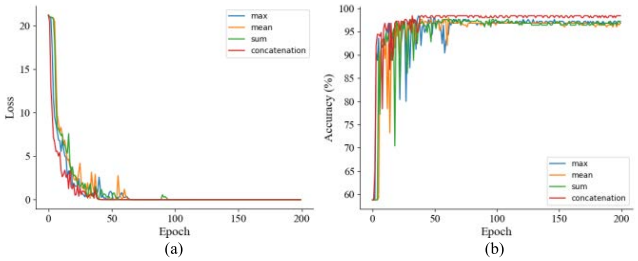

FIGURE 5. Training process. (a) Loss on training set. (b) Accuracy on validation set.

TABLE 6. The effects of each component in cross-attention.

Architecture	Accuracy rate (%)
Channel-wise attention	98.32(±0.33)
Feature interaction mechanism	98.40(±0)
Feature interaction mechanism & Channel-wise attention	98.24(±0.22)
Channel-wise attention & Feature interaction mechanism	98.64(±0.35)

E. NECESSITIES OF CROSS-ATTENTION MECHANISM

In previous experiments, it was obvious that using three branches and concatenation fusion method achieved the best results. On this basis, the cross-attention mechanism is introduced into MMFCNN to demonstrate the effectiveness. In addition, the order of components may have a great impact on the effect of MMFCNN. Therefore, experiments on the order of the exchange of components were carried out. Experimental results are shown in Table 6. It can be observed, using channel-wise attention followed by a feature interaction mechanism (cross-attention) can more effectively improve the performance of MMFCNN, whose accuracy rate is 98.64%. This is because channel-wise attention enables the network to focus on the information of internal defects, and the feature interaction mechanism enables the network to associate information between two branches during feature extraction. By contrast, the result of using a feature

interaction mechanism followed by channel-wise attention is slightly worse than the previous architecture. As a result of the locations of defect information in each modal data are different, and the features of one mode integrate the global features of another mode, resulting in the disorder of the channel-wise attention mechanism. However, using channel-wise attention or feature interaction mechanism alone cannot improve the accuracy of our proposed algorithm. From the training process, they can accelerate the convergence of the network and make the network more stable. Moreover, the confusion matrix of the validation set has also been given in Fig. 6. For failure cases of prediction, there are two reasons. On the one hand, long-hours working leads workers to mistakenly mark samples. On the other hand, these failure cases may exist tiny internal defects that are acoustically close to normal magnetic tiles.

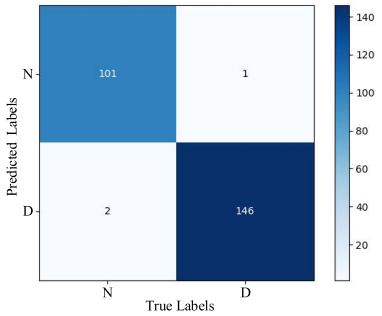

FIGURE 6. Confusion matrix of MMFCNN on validation set.

TABLE 7. Time complexity and inference time.

Architecture	FLOPs	Inference time(s)
MMFCNN	2.80G	0.08

F. FIELD VERIFICATION

Before the deployment of the model, we analyze the time complexity and inference time of MMFCNN. For time complexity, we calculate the FLOPs (floating-point operations per second) of proposed network. For inference time, we calculated the average value of the prediction time of 10 magnetic tile data. The specific results are shown in Table 7. To verify the adaptability of our proposed model to the whole detection system, we simulate the detection process of internal defects of magnetic tile, and build a similar and simple system. The NI-9250 sound acquisition card equipped with a sound sensor and the sound acquisition software system written by LabVIEW is used to sample the sound excited by magnetic tile and iron block in real-time. To avoid the influence of subjective factors, we test the newly produced magnetic tiles, and they are detected again by the experienced worker. Finally, the results of the two tests are compared. The comparison result is as follows. 100 newly produced magnetic tiles are tested through our established system. The test results show that 93 pieces are normal and 7 pieces have internal defects. The results of manual detection are consistent with ours. Therefore, this shows that our model has strong applicability.

G. INFLUENCE OF CONVOLUTIONAL PARAMETERS

Our backbone is based on AlexNet. The parameters in Table 1 are similar to AlexNet. For time domain and frequency domain data, because they are relatively sparse, small convolution kernels are used to extract neighborhood information. For time-frequency domain spectrogram, large convolution kernel is firstly used to extract a wide range of neighborhood information. Then, small convolution kernel is used to extract high-dimensional features. To demonstrate the influence of the different parameters, we mainly discuss the number of filters. The number of filters of five layers in our network is 1, 3, 6, 4 and 4 times that of the first layer. Keep the multiplier constant, and set the number of filters of the first layer as 32, 64, 96 and 128 to illustrate the influence of the number of filters on the network performance. And the corresponding parameters are marked as Conv1_X(32), Conv1_X(64), Conv1_X(96) and Conv1_X(128), respectively. The comparison results are shown in Table 8. As can be seen from Table 8, the number of filters in Table 1 can make the network achieve the highest accuracy.

TABLE 8. The influence of the number of filters.

The Number of filters	Accuracy rate (%)
Conv1_X(32)	98.0
Conv1_X(64) (MMFCNN)	98.8
Conv1_X(96)	98.4
Conv1_X(128)	98.4

V. CONCLUSION

In this work, a novel deep learning-based CNN named MMFCNN was proposed for the internal defect detection of magnetic tile. Based on this algorithm, a new intelligent system was developed, which can automatically obtain sound and identify the internal defects of the magnetic tile. To take advantage of multimodal information, we utilized multiple branches to extract features respectively, and then, carry out the feature fusion operation. And then, multiple feature fusion methods were discussed. Moreover, the cross-attention mechanism was constructed to realize the information interaction among branches and emphasize more representative features, which can improve the performance of our model. As for whether each module in the cross-attention mechanism is necessary, several ablation experiments were carried out. Extensive experimental results show that our model is superior for internal defect detection of magnetic tile.

In this article, we assume the training set and test set follow the same distribution. Besides, the number of defective samples is comparable to that of normal samples. However, the number of normal magnetic tile is much more than that of the defective in the production process. On the other hand, our model is only for one kind of magnetic tile, which is not

enough for the magnetic tile industry. Moreover, the proposed network is relatively large, which lead to the detection speed cannot be too high. Therefore, the related transfer learning, few-shot learning and knowledge distillation models need to be studied in our future work.

REFERENCES

[1] L. Xie, X. Xiang, H. Xu, L. Wang, L. Lin, and G. Yin, "FFCNN: A deep neural network for surface defect detection of magnetic tile," *IEEE Trans. Ind. Electron.*, vol. 68, no. 4, pp. 3506–3516, Apr. 2021, doi: 10.1109/TIE.2020.2982115.

[2] L. Xie, L. Lin, M. Yin, L. Meng, and G. Yin, "A novel surface defect inspection algorithm for magnetic tile," *Appl. Surf. Sci.*, vol. 375, pp. 118–126, Jul. 2016, doi: 10.1016/j.apsusc.2016.03.013.

[3] X. Li, H. Jiang, and G. Yin, "Detection of surface crack defects on ferrite magnetic tile," *NDT E Int.*, vol. 62, pp. 6–13, Mar. 2014, doi: 10.1016/j.ndteint.2013.10.006.

[4] G. Ji, "Research on nondestructive testing of microcracks," in *Proc. 2nd Int. Conf. Artif. Intell. Adv. Manuf. (AIAM)*, Oct. 2020, pp. 395–397, doi: 10.1109/AIAM50918.2020.00087.

[5] Q. Li, Q. Huang, Y. Zhou, T. Yang, M. Ran, and X. Liu, "Combined convolutional and LSTM recurrent neural networks for internal defect detection of arc magnets under strong noises and variable object types," *IEEE Access*, vol. 9, pp. 71446–71460, 2021, doi: 10.1109/ACCESS.2021.3078709.

[6] Q. Huang, Y. Yin, and G. Yin, "Automatic classification of magnetic tiles internal defects based on acoustic resonance analysis," *Mech. Syst. Signal Process.*, vols. 60–61, pp. 45–58, Aug. 2015, doi: 10.1016/j.ymssp.2015.02.018.

[7] L. Xie, Q. Huang, Y. Zhao, and G. Yin, "Inspection of magnetic tile internal cracks based on impact acoustics," *Nondestruct. Test. Eval.*, vol. 30, no. 2, pp. 147–164, Apr. 2015, doi: 10.1080/10589759.2015.1018255.

[8] L. Xie, M. Yin, Q. Huang, Y. Zhao, Z. Deng, Z. Xiang, and G. Yin, "Internal defect inspection in magnetic tile by using acoustic resonance technology," *J. Sound Vib.*, vol. 383, pp. 108–123, Nov. 2016, doi: 10.1016/j.jsv.2016.07.020.

[9] Q. Huang, L. Xie, G. Yin, M. Ran, X. Liu, and J. Zheng, "Acoustic signal analysis for detecting defects inside an arc magnet using a combination of variational mode decomposition and beetle antennae search," *ISA Trans.*, vol. 102, pp. 347–364, Jul. 2020, doi: 10.1016/j.isatra.2020.02.036.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, Sep. 2015, doi: 10.1038/nature14539.

[11] G. Hinton, "Where do features come from?" *Cognit. Sci.*, vol. 38, no. 6, pp. 1078–1101, Aug. 2014, doi: 10.1111/cogs.12049.

[12] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013, doi: 10.1109/TPAMI.2012.231.

[13] T. Han, C. Liu, W. Yang, and D. Jiang, "A novel adversarial learning framework in deep convolutional neural network for intelligent diagnosis of mechanical faults," *Knowl.-Based Syst.*, vol. 165, pp. 474–487, Feb. 2019, doi: 10.1016/j.knosys.2018.12.019.

[14] Y. He, K. Song, Q. Meng, and Y. Yan, "An end-to-end steel surface defect detection approach via fusing multiple hierarchical features," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 4, pp. 1493–1504, Apr. 2020, doi: 10.1109/TIM.2019.2915404.

[15] S. Xiaoping, C. Jinsheng, and G. Yuan, "A new deep learning method for underwater target recognition based on one-dimensional time-domain signals," in *Proc. OES China Ocean Acoust. (COA)*, Jul. 2021, pp. 1048–1051, doi: 10.1109/COA50123.2021.9520078.

[16] J. Yang, B. Zou, H. Qiu, and Z. Li, "MLFNet- point cloud semantic segmentation convolution network based on multi-scale feature fusion," *IEEE Access*, vol. 9, pp. 44950–44962, 2021, doi: 10.1109/ACCESS.2021.3057612.

[17] D. Wang, J. Su, and H. Yu, "Feature extraction and analysis of natural language processing for deep learning English language," *IEEE Access*, vol. 8, pp. 46335–46345, 2020, doi: 10.1109/ACCESS.2020.2974101.

[18] B. Zhang, D. Xiong, and J. Su, "Neural machine translation with deep attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 154–163, Jan. 2020, doi: 10.1109/TPAMI.2018.2876404.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, Jun. 2012.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[21] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," 2015, *arXiv:1505.00853*.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1, pp. 1–14, Apr. 2015.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

[24] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856, doi: 10.1109/CVPR.2018.00716.

[25] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, May 2019, pp. 10691–10700.

[26] R. Xiao, Z. Zhang, Y. Wu, P. Jiang, and J. Deng, "Multi-scale information fusion model for feature extraction of converter transformer vibration signal," *Measurement*, vol. 180, Aug. 2021, Art. no. 109555, doi: 10.1016/j.measurement.2021.109555.

[27] S. Zhang, X. Tao, Y. Chuang, and X. Zhao, "Learning deep multimodal affective features for spontaneous speech emotion recognition," *Speech Commun.*, vol. 127, pp. 73–81, Mar. 2021, doi: 10.1016/j.specom.2020.12.009.

[28] C. Ieracitano, N. Mammone, A. Hussain, and F. C. Morabito, "A novel multi-modal machine learning based approach for automatic classification of EEG recordings in dementia," *Neural Netw.*, vol. 123, pp. 176–190, Mar. 2020, doi: 10.1016/j.neunet.2019.12.006.

[29] Z. Ahmad, A. Tabassum, L. Guan, and N. M. Khan, "ECG heartbeat classification using multimodal fusion," *IEEE Access*, vol. 9, pp. 100615–100626, 2021, doi: 10.1109/ACCESS.2021.3097614.

[30] P. Qi, X. Zhou, S. Zheng, and Z. Li, "Automatic modulation classification based on deep residual networks with multimodal information," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 1, pp. 21–33, Mar. 2020, doi: 10.1109/TCCN.2020.3023145.

[31] M. Liang, P. Cao, and J. Tang, "Rolling bearing fault diagnosis based on feature fusion with parallel convolutional neural network," *Int. J. Adv. Manuf. Technol.*, vol. 112, nos. 3–4, pp. 819–831, Jan. 2021, doi: 10.1007/s00170-020-06401-8.

[32] B. Wang, Y. Lei, N. Li, and W. Wang, "Multiscale convolutional attention network for predicting remaining useful life of machinery," *IEEE Trans. Ind. Electron.*, vol. 68, no. 8, pp. 7496–7504, Aug. 2021, doi: 10.1109/TIE.2020.3003649.

[33] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, vol. 37, Jul. 2015, pp. 448–456.

[34] H. Xu, N. Ye, G. Liu, B. Zeng, and S. Liu, "FINet: Dual branches feature interaction for partial-to-partial point cloud registration," 2021, *arXiv:2106.03479*.

[35] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2017, doi: 10.1109/TPAMI.2019.2913372.

HOUHONG LU received the B.Sc. degree in mechanical engineering from the School of Mechanical Engineering, Shandong University, Jinan, China, in 2019. He is currently pursuing the M.S. degree in mechanical engineering with the School of Mechanical Engineering, Sichuan University, Chengdu, China. His current research interests include nondestructive testing and computer vision.

YANGYANG ZHU received the B.Sc. degree in mechanical engineering from the School of Mechanical Engineering, Hefei University of Technology, Hefei, China, in 2019. He is currently pursuing the M.S. degree in mechanical engineering with the School of Mechanical Engineering, Sichuan University, Chengdu, China. His current research interests include nondestructive testing, point cloud registration, and computer vision.

MING YIN received the B.Sc. and Ph.D. degrees from the School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, China, in 2010 and 2014, respectively. He is currently an Associate Professor with the School of Mechanical Engineering, Sichuan University, Chengdu, China. His current research interests include computer vision and data mining.

GUOFU YIN received the Ph.D. degree from the School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, China, in 1989. He is currently a Professor with the School of Mechanical Engineering, Sichuan University, Chengdu, China. He has published over 200 scientific papers. His current research interests include mechanical design and manufacturing, robotics and mechatronics, image processing and pattern recognition, computer vision, machine learning, and CAD/CAE/CAM.

LUOFENG XIE received the B.Sc. and Ph.D. degrees from the School of Manufacturing Science and Engineering, Sichuan University, Chengdu, China, in 2014 and 2019, respectively. He is currently an Assistant Professor with the School of Mechanical Engineering, Sichuan University. His current research interests include nondestructive testing, signal processing, and data mining.