IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 March 2023, accepted 19 May 2023, date of publication 29 May 2023, date of current version 12 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3280606

RESEARCH ARTICLE

The Classification and Judgment of Abnormal Problems in Music Song Interpretation Based on Deep Learning

ZHONGWEI XU[1], WEITE ZOU[2], YUAN FENG[1], SIQI LIU[2], YUANXIANG XU[1], SHENGYU SONG[1], LAN ZHANG[1], MIAOMIAO TIAN[1], AND JIAHAO LIU[1]

[1]College of Information Science and Engineering, Ocean University of China, Qingdao 266005, China
[2]Teaching Center of Fundamental Courses, Ocean University of China, Qingdao 266005, China

Corresponding author: Yuan Feng (fengyuan@ouc.edu.cn)

This work was supported in part by the National Research and Development Program of China under Grant 2020YFB1710401, and in part by the National Natural Science Foundation of China under Grant 61902367 and Grant 41976185.

ABSTRACT Song singing interpretation is one of the people's favorite entertainment pastimes, but there are some abnormal problems in the process of song singing. Traditional song singing problems need to be analyzed and judged by professionals, but the traditional way requires offline face-to-face teaching and is time-consuming and laborious. In this paper, we hope to realize the automatic judgment classification of abnormal vocal problems of song singing and provide some guidance help to online teaching. In this paper, we propose a deep learning-based method for classifying abnormal vocal interpretation problems in music songs, using a computer to record the singers' voices, and then analyzing and judging them with a trained method model to point out the main problems that exist in the process of singing songs. In this paper, more than 300 singers' audio were collected and the data were calibrated and classified by researchers specialized in the music field into seven main categories. Short-time Fourier transform (STFT), Mel frequency cepstrum coefficient (MFCC) and spectral mass center methods were used to extract the features of song audio and produce the corresponding datasets. The dataset is trained using residual neural network and EfficientNet. The experimental results of the model in this paper show that the data training accuracy is about 90.1%, which achieves a good result.

INDEX TERMS ResNet, EfficientNet, the spectral center of mass, deep learning, short time fourier transform, mel frequency cepstrum coefficient.

I. INTRODUCTION

Online teaching is a more popular way of teaching, some art training also realized online teaching, online teachers by listening to the sound coming out of the computer to determine the shortcomings of the students singing, but like the song singing teaching requirements are relatively high, online teaching there are many realistic problems, such as network delays, sound input, and output hardware equipment quality problems, etc., are affecting the teacher's judgment. So we use deep learning training to learn the audio recorded by the singer and realize the judgment and classification of the abnormal problems that existed in the song singer, which is a challenging work and has strong theoretical value and practical significance in the fields of self-learning and online song singing teaching for amateurs and abnormal problem detection.

In this paper, first of all, in the production of the data set, we collected more than 300 audios of the same song sung by different people. Since music singing has the artistry of personal characteristics, the main focus in determining the abnormal problems in the singing process is to compare with the original songs. We address the main problems in the singing process of singers and classify these audios into seven main categories of abnormal problems: pale vocalization (singing without emotion) Shuaiwu [1], Zhufeng [2], inability to keep up with the rhythm Hongzi [3], Jing [4], out-of-tune, insufficient breath Jianyang [5], Lin [6], unclear

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

spitting Maimedi [7], narrow range (not going up in the treble and not coming down in the bass) Shuaiwu [1], Kun [8], and true-falsetto voice problems. Manually, these datasets are calibrated and classified.

For the field of music classification, from the use of traditional classical classifiers, Kostrzewa et al. [9] used KNN classifiers to classify music by genre; Silla et al. [10] used machine learning to achieve the automatic classification of music genres. In the field of deep learning, Choi et al. [11] proposed a migration learning method for music classification and regression tasks; Lim et al. [12] proposed a convolutional neural network (CNN) based method for audio event classification. Choi et al. [13] used convolutional recurrent neural networks to classify music; Gunawan and Suhartono [14] used convolutional recurrent neural networks to implement a recommendation system for music genres; Wenkang [15] used a deep learning approach to classify music genres; Li and Bin [16] proposed a music content-based classification method that applied long and short-term memory networks (LSTM) in deep learning to classify music genres, combining different neural network approaches to explore music genre classification. Ahmad et al. [17] used hybrid convolutional-recursive neural network analysis techniques to analyze music genre classification. Kostrzewa et al. [18] used several deep learning models (CNN, CRNN, LSTM) for the genre classification of music tracks. Yi et al. [19] used MFCC as a popular acoustic feature to achieve the classification of music genres. Zhang et al. [20] used a convolutional recurrent neural network (C-RNN) to classify music. Liu et al. [21] implemented an unsupervised fault diagnosis method for rolling bearings using a short-time Fourier transform and generative neural network. Wang et al. [22] used improved CRNN with temporal and frequency multi-directional spatial dependence for music classification. Huan [23] based on the convolutional neural network using inverse spectral coefficients to extract MFCC features of audio. Youchen [24] proposed Dense Inception new convolutional neural network architecture for music genre classification. Ashraf [25] studied music classification using convolutional recurrent neural networks and residual learning. It is obvious from the above studies that with the rapid development and widespread use of artificial intelligence, deep learning models play an important role in music genre classification and audio anomaly detection discriminative classification. Deep learning also has relevant applications in areas such as anomaly detection and fault prediction, for example, Lee and Mitici [26] and Namdari et al. [27], in using deep reinforcement learning to achieve fault prediction for aircraft and lithium-ion batteries, respectively, reducing the related maintenance costs, yet improving the safety of the related equipment. So it is theoretically and technically feasible to use deep learning models to achieve the classification of song rendition abnormal vocalization problem determination.

Secondly, in terms of audio feature extraction, different problematic audio is bound to have some differences, and unique features are often considered a key step to achieving high accuracy in classification Zebari et al. [28]. We mainly identify the problems in the singing process of the singer, so we cut out the unnecessary unmanned voice frequencies such as the intro, and keep only the part with the human voice. We use Short Time Fourier Transform, MFCC, and spectral prime to achieve the extraction of this part of audio features, for Short Time Fourier Transform spectrograms are widely used for classification tasks Stowell et al. [29] and MFCC is a more popular feature extraction method for audio feature extraction Tirronen et al. [30], and also has a good application in the field of audio classification Deng et al. [31]. We have achieved better results using related feature extraction methods on music genre classification. In the processing of data, mostly the whole audio is feature extracted, which inevitably leads to data loss due to the limited size of trainable images. Based on the characteristics of music, the songs in this dataset are 2/4 beats, with each bar has 2 beats and each beat is about 2 seconds, so we intercepted in pieces according to the duration of each bar and superimposed the phases according to 2s per beat, which means that two adjacent segments have 50% of superimposed parts, so that we can achieve the maximum utilization of audio features and reduce the loss of data, while preserving the continuity and relevance of data.

For the selection of training models, we use the residual neural network (ResNet) and efficient neural network (EfficientNet), and we have achieved better results in music genre classification using residual neural network and efficient neural network in image classification Tan and Le [33]. We add spatial and channel attention mechanisms in training to effectively improve the training accuracy, and the best accuracy of the residual neural network is currently about 87.7%, and the training accuracy of the efficient neural network is 90.1%.

The main contributions of this paper are (1) the research of using deep learning techniques to realize the song rendition problem is proposed, and a large amount of learning training is conducted using deep learning to couple out the vocal abnormalities existing in the song rendition process, which is efficient compared to manual judgment. (2) In this study, targeted data pre-processing methods and feature extraction methods are selected to superimpose the data in pieces for the characteristics of music, enhance the continuity and relevance of data, and effectively improve the data volume and data feature utilization. And the extraction of audio features was realized by using STFT and MFCC, and the images for neural network training were generated, which effectively improved the effect of analysis and judgment of song interpretation problems. (3) The analysis and judgment of song rendition problem is realized by using residual neural network and EfficientNet, and the attention mechanism such as adding space and channel is optimized to improve the classification accuracy. As Table 1 shows the application scenarios and advantages and disadvantages of the neural network models mainly used in this paper.

TABLE 1. Introduction of the main neural network models used in this paper.

Model	Application Scenarios	Advantages	Disadvantages
ResNet-101	Image classification, target detection, image segmentation	Deep network structure that helps solve gradient disappearance and gradient explosion problems; can learn more complex feature representations; performs well on large-scale image datasets	Deeper models, longer training and inference time; requires more computational resources and storage space
EfficientNet	Image classification, target detection, image segmentation	Higher computational efficiency with limited computational resources; optimized at different scales with network scaling methods to improve performance; uses less memory and computational resources when deployed on devices with limited computational resources	Performance may be slightly reduced relative to larger models

II. MATERIALS AND METHODS

The dataset in this paper is made for neural network training requirements, and diverse dataset feature extraction methods are selected, combined with different deep learning models, and adjusted and optimized on this basis, with the ultimate goal of trying to use different feature extraction methods combined with different deep learning model frameworks to explore the best results for achieving anomaly classification of song renditions.

In this subsection, we focus on the dataset and hardware setup; then the method and process of audio feature extraction, to generate visual images for possible deep learning model training, and finally followed by an introduction to the optimization and architecture of the main deep learning models used in this paper.

The main research method of this paper used is the deep learning method, realizing the judgment and classification of abnormal vocal problems existing in the singing process of song singers. With the development of information technology and the rapid entry of mobile Internet into people's lives, many music APPs support personal music singing and sharing, and even singing teaching, but we found that the effect of singers' songs singing and the scoring results are not exactly equal. We have tried to find that, according to the same music song sung with different words, as long as the pitch can keep up with the scoring line given by the APP, the rating is generally not very bad, and it does not completely show one's singing level. In addition, traditional music singing teaching is mostly one-to-one instruction, which leads to high teaching costs and a low service population. So we resorted to using computerized deep learning to achieve an automated classification judgment of abnormal problems in song singers' singing, which could be a guide to online music teaching.

A. DATA SET AND HARDWARE SETUP

When we explored the use of deep learning methods to achieve judgmental classification of song rendition anomalies, we did not find a relevant publicly available dataset, so we teamed up with music professional researchers to help us produce a relevant dataset.

Song Anomaly Dataset (SAD) is a collection of audio of the same song sung by different people. We asked professional researchers to compare the audio of the song with the original song and classify the abnormal singing problems, which are divided into seven categories: singing pale without emotion, poor rhythm, lack of breath, unclear spitting, out of tune, narrow range and inaccurate switching between true and false voices. The SAD dataset contains a total of 300 tracks, with each 197 seconds long and each audio sample type is 44100Hz dual-channel.

TABLE 2. Data set categories and numbers.

Category	Quantity	Slice Quantity
Insufficient breath	50	4499
Narrow range	40	3599
Pale without emotion	50	4496
Rhythm	28	2520
Sing out of the tune	50	4421
True falsetto	50	4500
Unclear enunciation	50	4498

1) DATA PRE-PROCESSING

The model has a limit on the size of the images that can be input for training, and we process the audio according to the same image size, which inevitably leads to the compression of the unsliced audio features, resulting in a large loss of relevant data features. In addition, the number of data sets is too small, so it is difficult to achieve better results, and the experiments also prove that the training effect is bad with an accuracy rate of only about 30%. The piecewise overlay processing, which cuts the long audio into short audio, can show more information in the same size picture when feature extraction, maximizing data utilization, and the overlay processing can eliminate the loss of data edge features, which can effectively improve the training efficiency. As Table 2 shows the number of songs in each category of the dataset as well as related categories and the number of each category after binning.

The beat refers to the regular and periodic repetition of strong and weak beats of the same time value. It is the presentation of the musical structure and the lowest foundation of the musical composition. The song performer can effectively control the rhythm of the song (e.g., at what beat point the singer starts singing) through the beats to prevent problems in the process of singing. Therefore, slicing by beat duration can effectively obtain audio key node feature data. We based on the characteristics of music, this dataset song for 2/4 beats, each bar has 2 beats, and each beat about 2 seconds, so we slice interception according to the length of each bar, but because slicing will certainly lead to the loss of some edge

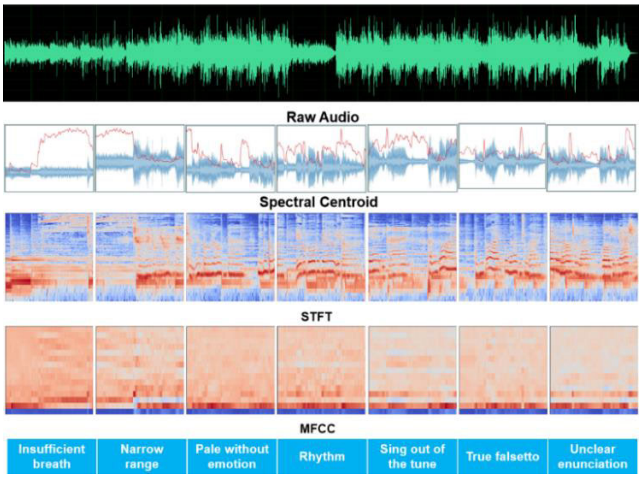

FIGURE 1. Extraction of features for each category of audio clips.

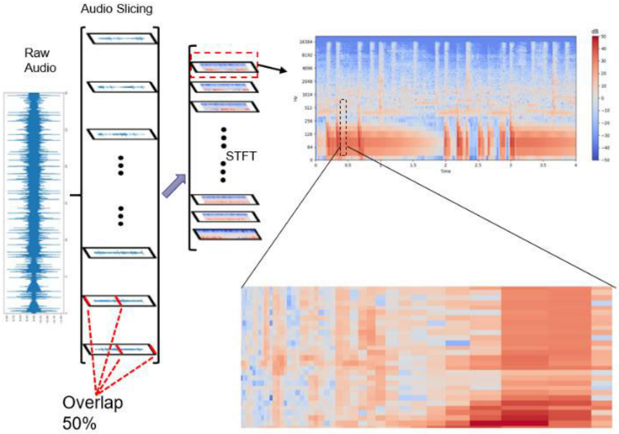

FIGURE 2. Short-time fourier transform audio feature extraction process.

data, and unsliced audio each audio independent relevance is not large, so and choose the superposition processing, according to 2s per beat for phase superposition, that is, two adjacent segments have 50% of the overlapping part. The 50% overlapping part between different clips can not only increase the data volume but also reduce the data loss and enhance the data features and data correlation, which also achieved better results in the experiment.

Because of the inconsistent frequency of anomalous problems in the song rendition process, the number of different categories is unbalanced when collecting and producing the dataset. In order to reduce the impact that the unbalanced dataset may bring to the final training results, this paper adjusts on the traditional cross-entropy loss function. N is the total number of datasets; $P_{n,j}$ is the probability that category j appears in the nth sample in the dataset.

$$\text{Loss} = -\frac{1}{N}\sum\nolimits_{n=1}^{N} log(P_{n,j}) \tag{1}$$

In this paper, we add the weight Y_n to the traditional cross-entropy function, if the number of samples with the largest amount of data in the dataset is X, and the number of other category data sets is T_i (i is other category data), so the weight data coefficient will be $Y_n = X/T_i$, and the category weight with the largest amount of data in the dataset here is 1. After such a weight is given, in the neural network training process, it will be selectively skewed to try to balance the overall training and learning.

$$\text{YLoss} = -\frac{1}{N}\sum\nolimits_{n=1}^{N} Y_n log(P_{n,j}) \tag{2}$$

2) FEATURE EXTRACTION

We mainly extracted three different features for each fragmented audio: Short Time Fourier Transform, Mel-frequency spectral coefficients, and spectral prime maps. The following figure shows the effect of feature extraction for each type of fragment of the original music audio and the three feature extraction methods. The data sets produced by the three feature extraction methods we named as SAD-Stft, SAD-Mfcc, and SAD-Sc. As shown in Figure 1.

B. FEATURE EXTRACTION

1) SHORT TIME FOURIER TRANSFORM (STFT)

Short-time Fourier transform has a relatively wide application in the field of audio feature extraction and can better reflect the features of different music genres Elbir et al. [34]. In Figure 2, the original audio is first binned (50% overlapping between adjacent segments), and the STFT feature extraction is performed on the binned audio to generate images.

The concrete steps of STFT: First of all, starting from the starting position of the audio signal, t = t_0 is the center of the window function, and we add window processing to the audio signal, and we set n_fft, the FFT window size, to 1024 here.

$$y(t) = x(t) \cdot \omega(t - t_0) \tag{3}$$

Fourier transform again:

$$X(\omega) = \mathcal{F}(y(t)) = \int_{\infty}^{+\infty} x(t) \cdot w(t - t_0) e^{-j\omega t} dt \tag{4}$$

The resulting spectral distribution $X(\omega)$ of the first segmented sequence is obtained, and after completing the FFT operation on the first segment, the window function is shifted to t_1 Shao et al. [32]. The distance the window is moved is referred to as the hop_ length: i.e., the frameshift, which we set to 512.

Repeat the above operation, keep sliding the window, FFT, and finally get the spectrum results of all segments from $t_0 \sim t_N$: finally, we get S, which is the result after STFT transformation.

2) MEL-FREQUENCY SPECTRAL COEFFICIENTS

MFCC is a further extension of the Mel spectrogram and Mel-Frequency cepstrum coefficients Gupta et al. [35]

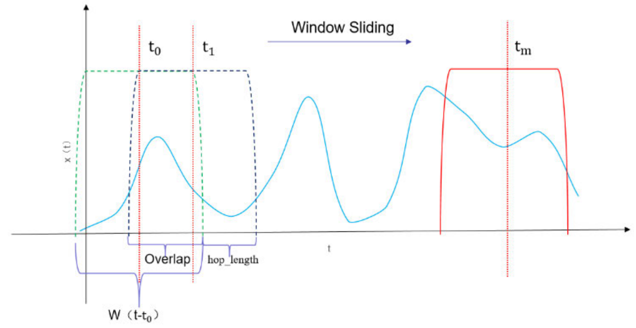

FIGURE 3. Schematic diagram of short-time fourier transform.

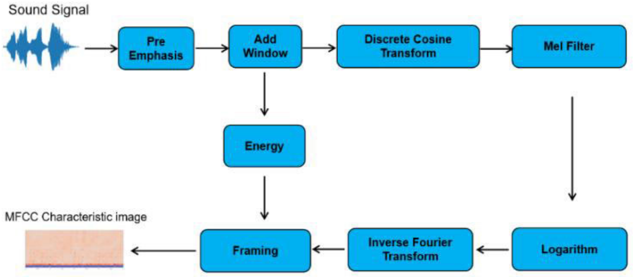
Sound Signal
Pre Emphasis
Add Window
Discrete Cosine Transform
Mel Filter
Energy
MFCC Characteristic image
Framing
Inverse Fourier Transform
Logarithm

FIGURE 4. MFCC feature extraction process diagram.

is another representative method after audio clip spectrum compression. The original audio is first binned first (50% overlapping between adjacent clips), and the feature extraction of the binned audio is performed to generate the image. As Figure 4 shows the MFCC feature extraction process diagram, firstly, the initial audio signal is pre-emphasized to compensate for the high-frequency audio signal to improve the recognition rate; then the window is added, the window size is set to 1024, and the frameshift is set to 512; then the discrete Fourier transform Gonzalez [36] is performed to map the audio signal from the time domain to the frequency domain; then the Meier filtering is performed and then the logarithm; then inverse Fourier transform, i.e., cepstrum analysis; differential selection of the first 14-dimensional cepstrum coefficients and energy for the 15th-dimensional features.

3) SPECTROSCOPIC MASS SPECTROMETRY

The spectral centroid is one of the important physical parameters to describe the timbral properties and is the centroid of the spectral distribution Han et al. [37], which is the center of mass of the signal frequency weighted by energy and reflects the objective distribution of sound energy in Hz. It is important information about the frequency distribution and energy distribution of the sound signal. The spectral center of mass has a good rationale for audio feature extraction in the field of music classification Sihan [38]. In the field of subjective perception, the spectral center of mass describes the brightness of a sound; sounds with a gloomy, low quality tend to have more low-frequency content and a relatively low spectral center of mass, and those with a bright, cheerful quality mostly focus on high frequencies and have a relatively high spectral center of mass Huang et al. [39]. This parameter is commonly used in the analytical study of the sound color of musical instruments. As the following equation where SC (Hz) is the spectral center of mass, f maximum and f minimum are the limit values of the signal frequency range, $E(f)$ is the corresponding frequency domain energy spectrum of the continuous time domain signal transformed by the short time Fourier transform, and f is the corresponding frequency $E(f)$ Xu et al. [40].

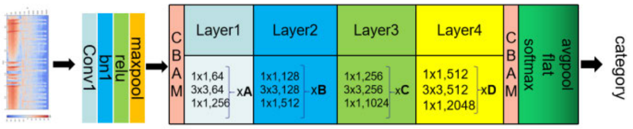

FIGURE 5. Resnet network structure diagram.

$$SC = \frac{\int_{f_{min}}^{f_{max}} fE(f)\, df}{\int_{f_{min}}^{f_{max}} E(f)\, df} \tag{5}$$

C. NEURAL NETWORK MODEL

1) RESIDUAL NEURAL NETWORK

The network architecture of the Resnet model with the addition of the over-convolutional block attention module (CBAM) is shown in Figure 5. CBAM is a simple and effective attention module that is mainly used for feedforward convolutional neural networks and is a lightweight general-purpose attention module Woo et al. [41]. The Resnet residual neural network protects information integrity by directly bypassing the input information to the output that protects the integrity of the information, and the whole network only needs to learn the part of the input and output difference, simplifying the learning goal and difficulty, which can be stacked can constitute a very deep network and has a high training efficiency. We also implemented adding channels and spatial attention mechanisms to the network to improve the model training accuracy He et al. [42]. In the experiments one can input images generated by MFCC, STFT, or spectral prime feature extraction, first by a convolutional layer stride of 2, then by a 3 × 3 maximal downsampling layer stride of 2, then a CBAM module containing a spatial attention mechanism and a channel attention mechanism, immediately followed by four residual blocks, then another CBAM module connected to the same. Finally, the average downsampling layer and the fully connected layer are used as outputs, along with softmax processing, to achieve training classification of audio categories.

The following table 3 shows the settings of network data parameters and related adjustment parameters for the resnet-101+CBAM model. In this experiment, the input image size is imageinput: 224 × 224 × 3; the number of iterations is epochs = 150; the size of each batch called batch_size = 16; the selection of optimizer and related data settings Adam:lr = 0.0001, these are the relevant parameters settings for the best training results.

TABLE 3. Resnet-101+cbam network structure and related parameter settings.

Layer Name	Layer Type	Output Shape	# of Parameters
conv1	Conv2d	(batch, 64, 112, 112)	9408
bn1	BatchNorm2d	(batch, 64, 112, 112)	128
relu	ReLU	(batch, 64, 112, 112)	0
ca	ChannelAttention	(batch, 64, h/2, w/2)	266496
sa	SpatialAttention	(batch, 64, h/2, w/2)	313
maxpool	MaxPool2d	(batch, 64, 56, 56)	0
layer1	Bottleneck	(batch, 256, 56, 56)x3	215808
layer2	Bottleneck	(batch, 512, 28, 28)x4	1219584
layer3	Bottleneck	(batch, 1024, 14, 14)x23	26111232
layer4	Bottleneck	(batch, 2048, 7, 7)x3	14964992
ca	ChannelAttention	(batch, 64, h/2, w/2)	266496
sa	SpatialAttention	(batch, 64, h/2, w/2)	313
avgpool	AdaptiveAvgPool2d	(batch, 2048, 1, 1)	0
fc	Linear	(batch, 1000)	2049000
Relevant parameter settings： imageinput:224×224×3； epochs=150； batch_size=16;Adam： lr=0.0001			

TABLE 4. EfficientNet_B7 network structure and related parameter settings.

Layer	Output Shape	Number of Parameters
Stem	(batch_size, 256, 256, 3)	5424
Block1	(batch_size, 128, 128, 48)	59661
Block2	(batch_size, 64, 64, 32)	123297
Block3	(batch_size, 32, 32, 56)	444000
Block4	(batch_size, 16, 16, 112)	2131184
Block5	(batch_size, 8, 8, 224)	7643808
Block6	(batch_size, 4, 4, 448)	29456576
Block7	(batch_size, 2, 2, 896)	119389696
Top	(batch_size, 1, 1, 3072)	52267072
Dropout	(batch_size, 1, 1, 3072)	0
FC	(batch_size, num_classes)	3073 * num_classes
Relevant parameter settings： imageinput: 600×600×3； epochs=150； batch_size=8; Adam： lr=0.0001,weight_decay=1E-4		

2) EFFICIENTNET

EfficientNet has a good application in the field of image classification Ase et al. [43] and is one of the more advanced models in the ImageNet classification problem. Most of the neural network optimization enhancements are done in three ways: first, increasing the number of convolutional kernels; second, increasing the depth of the neural network and building deeper and more layer structures; and third, improving the image resolution of the input neural network. The EfficientNet network is optimized to improve the performance of the network from all three perspectives simultaneously, as shown below for the baseline network and the compound scaling method, it is obvious to see that the compound scaling increases both the number of convolutional kernels and layers of the network depth compared to baseline, and there is a further improvement of input image resolution has been further improved Tan and Le [33], that is, EfficientNet achieves more efficient results by scaling the depth, width, and resolution evenly while shrinking the model Atila et al. [44]

The following table 4 shows the EfficientNet_B7 network structure and related parameters settings. In this experiment, the image input size is imageinput: 600 × 600 × 3; the number of iterations is epochs = 150; the data size of each batch call is batch_size = 8; the selection of the optimizer and related parameters are set as Adam:lr = 0.0001, weight_decay = 1E-4. These are the relevant parameter settings for the best training results.

All experiments are conducted on a desktop computer with the following hardware configurations: CPU-Intel Core (TM) i7-9700K, RAM-32GB RAM, GPU-NVIDIA GeForce GTX 2080 SUPER 8GB GDDR5.

III. RESULTS

Initially, we planned to perform feature extraction of the whole music directly without processing the original audio and tried STFT for feature extraction of music to produce a dataset for deep learning training. However, many problems were found, firstly, the amount of data was too small; secondly, the data utilization was low, the loss was large, and the image size that could be learned by the deep learning model was limited; there was also unnecessary audio data interference; finally, the training accuracy was extremely low and the loss could not be lowered.

We made targeted improvements for the problems that emerged in the experiment. During the song-singing process, not every lyric had abnormal problems, but overall each fragment had some of the same problems, plus the music had a beat rhythm, so we sliced the music according to the bars, so that the original data became close to 50 data, and the amount of data increased by 50 times, but it was found that there would be loss of data edge in the cropping process, and it led to little correlation, so we realized the superposition of music fragments according to the length of the bars to enhance the continuity and correlation of the data. We also remove some of the audio from unvoiced singing to reduce irrelevant interference. Then finally, according to the characteristics of the deep learning model, we set specific size images during feature extraction to improve the utilization of data features to enhance the training effect. The final dataset we produced is nearly 90 times larger than the initial experiment compared to the data volume, which effectively supports the subsequent experiments.

A. QUANTITATIVE RESULTS

The data is divided into three parts: training data, validation data, and test data, which are divided into 80% training data, 10% validation data, and 10% test data. The data pieces can be called randomly for training.

The following figure 7 shows the training accuracy of calling the SAD-Stft dataset and using various models. The lowest accuracy of the AlexNet neural network is 77.4%, and the highest accuracy is using EfficientNet_b7 neural network dataset image size 600 × 600 size accuracy, which is 90.1% as shown in the red part of the figure, and the overall accuracy is above 77%. In contrast to EfficientNet_b7, using the image size of 224 × 224 to 600 × 600, there is a significant improvement in the training accuracy, which shows that the

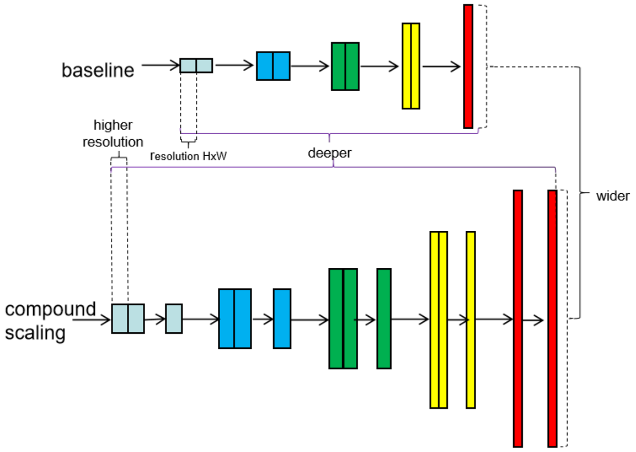

FIGURE 6. Baseline network and compound scaling method.

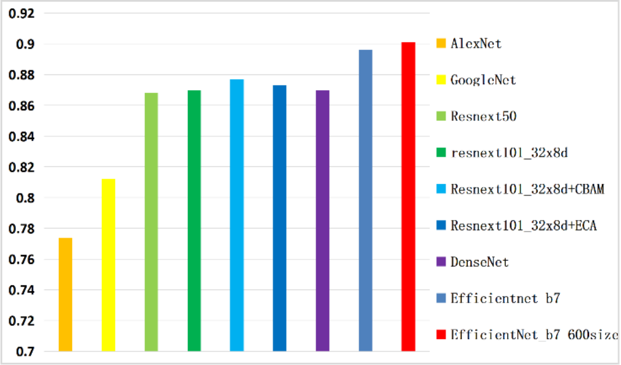

FIGURE 7. STFT features training accuracy of various models.

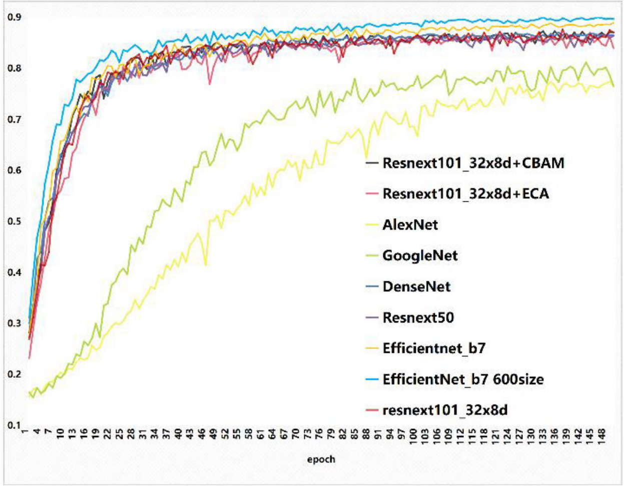

FIGURE 8. Model training accuracy comparison chart.

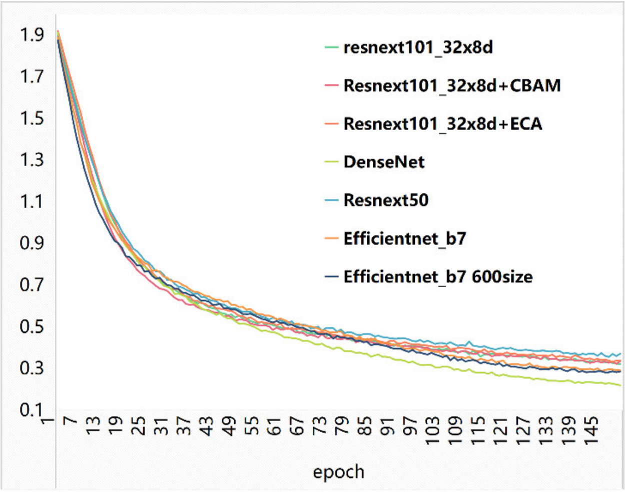

FIGURE 9. Model loss comparison chart.

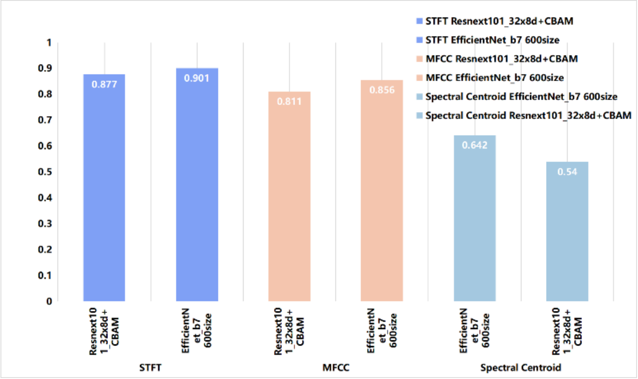

FIGURE 10. Training accuracy of different feature extraction methods.

feature image size has a greater impact on neural network training.

The following figure shows the model training accuracy comparison chart. We choose the number of iterations 150 times, compared to AlexNet and GoogleNet neural networks, other neural networks in the iteration to about 30 times accuracy has reached about 80% training accuracy improvement is relatively fast, and the latter is more stable. EfficientNet neural network in the late improvement is relatively large improvement close to 10%.

The following figure shows the comparison of the training model loss. Except for the EfficientNet neural network, other neural network gradient decline relatively fast, and the DenseNet network decline trend is relatively good, but finally, each network's final loss is down to about 0.3.

After the training comparison test using various types of neural networks on the dataset made by the STFT feature extraction method, we found that the residual neural network plus CBAM and efficient neural network (Resnext + CBAM and EfficientNet) can achieve better results than other types of neural networks in comparison, so these two types of neural networks were mainly used when training with SAD-Mfcc and SAD-Sc datasets.

As shown in Figure 10, the accuracy of the models trained by SAD-Stft, SAD-Mfcc, and SAD-Sc data using Resnext101_32x8d+CBAM and EfficientNet_b7 neural networks, respectively, can be seen from the figure that the STFT feature image training effect is relatively good accuracy

TABLE 5. Prediction accuracy of the validation set fragments.

Category	Insufficient breath	Narrow range	Pale without emotion	rhythm	Sing out of the tune	True falsetto	Unclear enunciation
Accuracy	88.9%	91.4%	93.6%	89.6%	93.1%	92.9%	97.6%

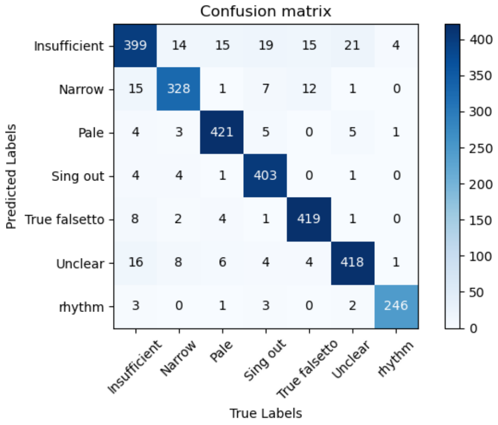

FIGURE 11. EfficientNet_b7 validation confusion matrix.

is the highest, followed by MFCC also reached 80% or more, while the best result of Spectral Centroid is only 64.2%.

The main reason is that the spectral mass method is more friendly to linear data and less effective to nonlinear data, and the spectral mass method is also more susceptible to noise, so it is less effective.

B. QUALITATIVE RESULTS

The following figure shows the best model confusion matrix using the EfficientNet_b7 neural network under STFT feature extraction, due to the difference in data volume, not every category validation set number is the same, only rhythm and Narrow number is less 252 and 359 respectively, but as shown in the figure, the higher prediction accuracy of each category is concentrated in the diagonal line.

As shown in Table 5 are the prediction accuracies of the above validation set segments. Among them, the accuracy of Unclear enunciation reached 97.6, and the accuracy of all categories was above 90% except Insufficient breath and rhythm which were close to 90% achieving better results.

The following table 6 shows the Precision, Recall, Specificity, and F1 scores for each category. Insufficient breath is relatively low in each category, but the performance indicators are above 80%, while the performance indicators of the remaining categories are above 90%, especially the Unclear enunciation category is above 96%.

Figure 12 shows the flow chart of song interpretation problem analysis and judgment. Firstly, the submitted audio is sliced and overlaid and filtered out the audio that is not related to the human voice, and then the feature extraction is performed on the sliced audio to generate the relevant feature extraction image, and then the image is input into the trained deep learning model, and the model performs automatic analysis to determine whether there is a relevant abnormal vocal problem in the segment and gives the problem category and its related accuracy.

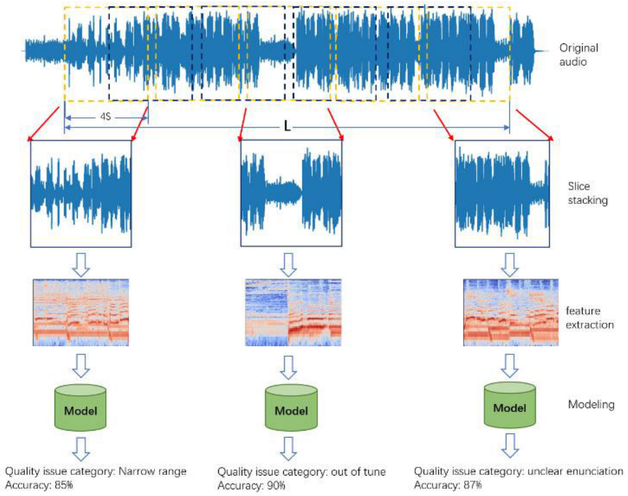
Original audio
4S
L
Slice stacking
feature extraction
Model
Model
Model
Modeling
Quality issue category: Narrow range
Accuracy: 85%
Quality issue category: out of tune
Accuracy: 90%
Quality issue category: unclear enunciation
Accuracy: 87%

FIGURE 12. Analysis and evaluation flow chart.

TABLE 6. EfficientNet training effect.

Category	Precision	Recall	Specificity	F1-Score
Insufficient breath	0.819	0.889	0.963	0.853
Narrow range	0.901	0.914	0.986	0.907
Pale without emotion	0.959	0.938	0.993	0.948
rhythm	0.976	0.912	0.996	0.943
Sing out of the tune	0.963	0.931	0.993	0.947
True falsetto	0.915	0.931	0.984	0.923
Unclear enunciation	0.965	0.976	0.997	0.970

IV. DISCUSSION

In this study, the judgment classification of abnormal vocal problems existing in the process of music singing was implemented, the audio sung by nearly 500 people was collected and professionally calibrated and classified by professional music researchers, and a rich and effective dataset was produced by using three ways of audio feature extraction: STFT, MFCC, and spectral mass center. We also used a variety of neural networks for comparative training tests to establish the best training model, and successfully achieved the efficient judgment and classification of the song-singing vocal anomaly problem. The most important thing in this paper is that we achieved the highest accuracy rate of 90.1%.

An advantage of this paper is the novel perspective and the richness of the data set. Music classification is a relatively hot field, but mostly the classification of music genres, we propose a classification for the determination of abnormal vocal problems during song interpretation, which can provide a reliable reference for distance music singing teaching. The dataset is the singing of nearly 300 different people, and the dataset is rich and diverse with different level dimensions, which avoids the interference of homogeneous singing of the same people, and such a dataset is more reliable.

Another advantage of this paper is the method of data processing, which reduces the interference of irrelevant information, enhances the continuity and relevance of data, and effectively improves the amount of data and data utilization. For neural networks, the amount of training data affects the effectiveness and accuracy of learning, and the size of the image classification image and the number of features presented determine the final effectiveness. We firstly filter out the unmanned part to reduce interference; secondly, slice processing to increase the amount of data; thirdly, superposition processing according to the characteristics of music beats to reduce data loss and enhance data continuity and correlation; fourthly, random selection during training.

Another advantage of this paper is the variety of feature extraction methods and matching model requirements. Three feature extraction methods are used, and again the parameters related to the three feature extraction methods are tested and adjusted, and matching image sizes are made for different model demand sizes to reduce data loss.

Another advantage of this paper is the diverse neural networks. We try to use Resnet50, Resnet101_32x8d, Densenet, Efficientnet_b7, Alexnet, and Googlenet a variety of networks for training tests on data sets with different characteristics, and Resnet101_32x8d network is adjusted to optimize the addition of attention mechanism to effectively improve the accuracy of training, and finally achieve the classification of the best network model, and achieve a high accuracy rate.

V. CONCLUSION

This section is not mandatory but can be added to the manuscript if the discussion is unusually long or complex.

The purpose of this paper is to provide a new and effective classification judgment method for singing vocal anomalies in the field of music singing teaching, to help music pedagogues or self-learning learners to assist in teaching or self-correction for improvement. In this study, we investigate from the perspective of song-singing vocal abnormalities and innovatively propose a classification judgment for this type of problem. We collected relevant data extensively and had music research professionals spend a lot of time on calibration, and searched for a suitable dataset production method in the process of continuous experimental exploration. We slice and superimpose the data according to the characteristics of music audio and according to the length of the bars, which reduces the data loss caused by generating images of the whole music, and at the same time enhances the utilization rate of the data and improves the different segments between continuity and correlation; and filtering out the drone part to avoid the interference of unnecessary data and improve the recognition rate. The STFT, MFCC, and spectral prime feature extraction methods are used to produce feature images that can be used for neural network learning, and the relevant parameters are adjusted to improve the feature extraction effect. And use Resnet50, Resnet101_32x8d, Densenet, Efficientnet_b7, Alexnet, Googlenet, and other neural networks to adjust the test to find the suitable neural network model, and through a large amount of data, training to generate and retain the best training model of Efficientnet_b7 and finally use the best training model for prediction judgment classification to achieve better prediction results. In the future, we will work to achieve real-time monitoring of abnormal problems in the singing process.

REFERENCES

[1] H. Shuaiwu, "Common mistakes in vocalization and correction for beginners," *Popular Songs*, vol. 10X, no. 10, p. 1, 2016.

[2] W. Zhufeng, "How to identify and correct wrong vocalizations in singing learning," *Contemp. Music*, vol. 16, no. 8, p. 2, 2016.

[3] Z. Hongzi, "Misunderstandings of singing and vocalization and training methods to correct them," *J. Chuzhou Univ.*, vol. 16, no. 3, p. 4, 2014.

[4] S. Jing, "Research on vocal singing skills and vocal methods," *Artists*, vol. 7, no. 7, p. 1, 2018.

[5] G. Jianyang, "Re-understanding of the scientificity of the method ofsinging and vocalization-discriminating and analyzing misunderstandings and moving towards scientific vocalization," *Art Educ.*, vol. 199, no. 2, p. 2, 2010.

[6] C. Lin, "Common misunderstandings and teaching countermeasures insinging and vocalization," *Music Grand View*, vol. 370, no. 2, p. 1, 2014.

[7] G. Maimedi, "Vocal singing skills and vocalization methods," *Charming China*. vol. 38, no. 1, p. 35, 2020.

[8] L. Kun, "Analysis of common mistakes and countermeasures in the learning of vocal music and singing skills," *Northern Music*, vol. 39, no. 3, p. 2, 2019.

[9] D. Kostrzewa, R. Brzeski, and M. Kubanski, "The classification of music by the genre using the KNN classifier," in *Proc. BDAS*, in Communications in Computer and Information Science, S. Kozielski, D. Mrozek, P. Kasprowski, B. Malysiak-Mrozek, and D. Kostrzewa, Eds., vol. 928. Cham, Switzerland: Springer, 2018, pp. 233–242, doi: 10.1007/978-3-319-99987-6_18.

[10] C. N. Silla Jr., A. L. Koerich, and C. A. A. Kaestner, "A machine learning approach to automatic music genre classification," *J. Brazilian Comput. Soc.*, vol. 14, no. 3, pp. 7–18, Sep. 2008.

[11] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Transfer learning for music classification and regression tasks," 2017, *arXiv:1703.09179*.

[12] M. Lim, D. Lee, H. Park, Y. Kang, J. Oh, J.-S. Park, G.-J. Jang, and J.-H. Kim, "Convolutional neural network based audio event classification," *KSII Trans. Internet Inf. Syst.*, vol. 12, no. 6, pp. 2748–2760, 2018.

[13] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2392–2396.

[14] Adiyansjah, A. A. S. Gunawan, and D. Suhartono, "Music recommender system based on genre using convolutional recurrent neural networks," *Proc. Comput. Sci.*, vol. 157, pp. 99–109, Jan. 2019.

[15] L. Wenkang, "Research on music genre classification based on the deep neural network," M.S. dissertation, South China Univ. Technol, Guangzhou, China, 2017.

[16] H. Li and Y. Bin, "Using long short-term memory network to classify music genres," *Comput. Technol. Develop.*, vol. 29, no. 11, p. 5, 2019.

[17] F. Ahmad and Sahil, "Music genre classification using spectral analysis techniques with hybrid convolution-recurrent neural network," *Int. J. Innov. Technol. Exploring Eng.*, vol. 9, no. 1, pp. 149–154, Nov. 2019.

[18] D. Kostrzewa, P. Kaminski, and R. Brzeski, "Music genre classification: Looking for the perfect network," in *Computational Science—ICCS 2021* (Lecture Notes in Computer Science), vol. 12742, M. Paszynski, D. Kranzlmüller, V. V. Krzhizhanovskaya, J. J. Dongarra, and P. M. A. Sloot, Eds. Cham, Switzerland: Springer, 2021, doi: 10.1007/978-3-030-77961-0_6.

[19] Y. Yi, K.-Y. Chen, and H.-Y. Gu, "Mixture of CNN experts from multiple acoustic feature domain for music genre classification," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2019, pp. 1250–1255.

[20] C. Zhang, Y. Zhang, and C. Chen, *SongNet: Real-Time Music Classifification*. Palo Alto, CA, USA: Stanford University Press, 2019.

[21] C. Liu, L. Feng, G. Liu, H. Wang, and S. Liu, "Bottom-up broadcast neural network for music genre classification," *Multimedia Tools Appl.*, vol. 80, pp. 7313–7331, Oct. 2019.

[22] Z. Wang, S. Muknahallipatna, M. Fan, A. Okray, and C. Lan, "Music classification using an improved CRNN with multi-directional spatial dependencies in both time and frequency dimensions," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[23] L. Huan, "Music genre classification based on convolutional neural network," *Electron. Meas. Technol.*, vol. 42, no. 21, p. 4, 2019.

[24] D. Youchen, "Research on music genre classification based on convolutional neural network," M.S. dissertation, Dalian Univ. Technol., Dalian, China, 2019.

[25] M. Ashraf, "Research on music classification algorithm based on convolutional recurrent neural network and residual learning," Ph.D. dissertation, Northwest Univ, Xi'an, China, 2021.

[26] J. Lee and M. Mitici, "Deep reinforcement learning for predictive aircraft maintenance using probabilistic remaining-useful-life prognostics," *Rel. Eng. Syst. Saf.*, vol. 230, Feb. 2023, Art. no. 108908.

[27] A. Namdari, M. A. Samani, and T. S. Durrani, "Lithium-ion battery prognostics through reinforcement learning based on entropy measures," *Algorithms*, vol. 15, no. 11, p. 393, Oct. 2022.

[28] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction," *J. Appl. Sci. Technol. Trends*, vol. 1, no. 2, pp. 56–70, May 2020.

[29] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1733–1746, Oct. 2015.

[30] S. Tirronen, S. R. Kadiri, and P. Alku, "The effect of the MFCC frame length in automatic voice pathology detection," *J. Voice*, Apr. 2022.

[31] M. Deng, T. Meng, J. Cao, S. Wang, J. Zhang, and H. Fan, "Heart sound classification based on improved MFCC features and convolutional recurrent neural networks," *Neural Netw.*, vol. 130, pp. 22–32, Oct. 2020.

[32] H. Shao, H. Jiang, H. Zhang, W. Duan, T. Liang, and S. Wu, "Rolling bearing fault feature learning using improved convolutional deep belief network with compressed sensing," *Mech. Syst. Signal Process.*, vol. 100, pp. 743–765, Feb. 2018.

[33] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[34] A. Elbir, H. O. Ilhan, G. Serbes, and N. Aydin, "Short time Fourier Transform based music genre classification," in *Proc. Electr. Electron., Comput. Sci., Biomed. Eng. Meeting (EBBT)*, 2018, pp. 1–4.

[35] S. Gupta, J. Jaafar, W. F. W. Ahmad, and A. Bansal, "Feature extraction using MFCC," *Signal Image Process. Int. J.*, vol. 4, no. 4, pp. 101–108, 2013.

[36] J. S. Jin et al., "Better than MFCC audio classification features," in *The Era of Interactive Media*. New York, NY, USA: Springer, 2013.

[37] N. C. Han, S. V. Muniandy, and J. Dayou, "Acoustic classification of Australian anurans based on hybrid spectral-entropy approach," *Appl. Acoust.*, vol. 72, no. 9, pp. 639–645, Sep. 2011.

[38] F. Sihan, "Music genre classification based on improved BP neural network," *Sofiw. Eng.*, vol. 24, no. 9, pp. 17–20, 2021.

[39] C.-J. Huang, Y.-J. Yang, D.-X. Yang, and Y.-J. Chen, "Frog classification using machine learning techniques," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3737–3743, Mar. 2009.

[40] X. Xu, J. Cai, N. Yu, Y. Yang, and X. Li, "Effect of loudness and spectral centroid on the music masking of low frequency noise from road traffic," *Appl. Acoust.*, vol. 166, Sep. 2020, Art. no. 107343.

[41] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, *CBAM: Convolutional Block Attention Module*. Cham, Switzerland: Springer, 2018.

[42] 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.

[43] A. S. Ebenezer et al., "Effect of image transformation on efficient net model for COVID-19 CT image classification," *Mater. Today, Proc.*, vol. 51, no. 8, pp. 2512–2519, 2022.

[44] Ü. Atila, M. Uçar, K. Akyol, and E. Uçar, "Plant leaf disease classification using EfficientNet deep learning model," *Ecol. Informat.*, vol. 61, Mar. 2021, Art. no. 101182.

ZHONGWEI XU received the bachelor's degree in engineering from Zhoukou Normal University, Henan, China, in 2020. He is currently pursuing the master's degree in computer technology with the Department of Information Science and Engineering, School of Computer Science and Technology, Ocean University of China.

WEITE ZOU received the bachelor's degree from the Shandong Academy of Arts, the master's degree from the Queensland Conservatory of Music, Australia, and the Ph.D. degree from the Shanghai Conservatory of Music. He is currently a Professor with the Art Department, Ocean University of China, and the Master's Supervisor and the Director of the Vocal Music Teaching and Research Office. He holds two patents. His research interests include diagnostics, plasma propulsion, and innovation plasma applications. He served as the Deputy Chairperson for the Branch of the Ocean University of China in the Democratic Revolution, the Secretary-General for the Laoshan District Musicians Association in Qingdao, and the Chairperson for the Ninth Postgraduate Meeting of Shanghai Conservatory of Music. He is an Associate Editor of the journal *Earth, Moon, and Planets*.

YUAN FENG was born in China, in 1978. He received the Ph.D. degree from the Ocean University of China, Qingdao, China, in 2008. He is currently a Professor with the Department of Information Science and Engineering, Ocean University of China. His research interests include sensor networks and cloud computing.

SIQI LIU received the bachelor's degree in music from the Ocean University of China, in 2020, where she is currently pursuing the master's degree in music literature.

YUANXIANG XU received the bachelor's degree in neo-confucianism from Qingdao University, Qingdao, Shandong, China, in 2021. He is currently pursuing the master's degree with the Ocean University of China, Qingdao.

SHENGYU SONG received the bachelor's degree in engineering from Qufu Normal University, Shandong, China, in 2020. She is currently pursuing the master's degree with the Ocean University of China.

LAN ZHANG received the bachelor's degree in engineering from Hefei University, Hefei, Anhui, China, in 2020. She is currently pursuing the M.Eng. degree with the Ocean University of China, Qingdao, Shandong, China.

MIAOMIAO TIAN received the bachelor's degree in engineering from the Anyang Institute of Technology, Anyang, Henan, China, in 2019. She is currently pursuing the master's degree in agronomy with the Ocean University of China, Qingdao, Shandong, China.

JIAHAO LIU received the bachelor's degree in engineering from Qingdao Agricultural University, Yantai, Shandong, China, in 2020. He is currently pursuing the M.Eng. degree with the Ocean University of China, Qingdao, Shandong.

• • •