Article

The Sustainable Development of Intangible Cultural Heritage with AI: Cantonese Opera Singing Genre Classification Based on CoGCNet Model in China

Qiao Chen ¹, Wenfeng Zhao ^{1,2,*}, Qin Wang ³ and Yawen Zhao ¹

¹ College of Electronic Engineering (College of Artificial Intelligence), South China Agricultural University, Guangzhou 510642, China; chen42ty@163.com (Q.C.); w844718456@163.com (Y.Z.)

² South China Smart Agriculture Public R&D (Research & Development) Platform, Ministry of Agriculture and Rural Affairs, Guangzhou 510520, China

³ Guangdong Art Research Institute, Guangzhou 510075, China; wqinr99@163.com

* Correspondence: zwf555@scau.edu.cn; Tel.: +86-1353-899-8771

Abstract: Chinese Cantonese opera, a UNESCO Intangible Cultural Heritage (ICH) of Humanity, has faced a series of development problems due to diversified entertainment and emerging cultures. While, the management on Cantonese opera data in a scientific manner is conducive to the sustainable development of ICH. Therefore, in this study, a scientific and standardized audio database dedicated to Cantonese opera is established, and a classification method for Cantonese opera singing genres based on the Cantonese opera Genre Classification Networks (CoGCNet) model is proposed given the similarity of the rhythm characteristics of different Cantonese opera singing genres. The original signal of Cantonese opera singing is pre-processed to obtain the Mel-Frequency Cepstrum as the input of the model. The cascade fusion CNN combines each segment's shallow and deep features; the double-layer LSTM and CNN hybrid network enhance the contextual relevance between signals. This achieves intelligent classification management of Cantonese opera data, meanwhile effectively solving the problem that existing methods are difficult to classify accurately. Experimental results on the customized Cantonese opera dataset show that the method has high classification accuracy with 95.69% Precision, 95.58% Recall and 95.60% *F1* value, and the overall performance is better than that of the commonly used neural network models. In addition, this method also provides a new feasible idea for the sustainable development of the study on the singing characteristics of the Cantonese opera genres.

Keywords: intangible cultural heritage; sustainable development; Cantonese opera; artificial intelligence; deep learning

Citation: Chen, Q.; Zhao, W.; Wang, Q.; Zhao, Y. The Sustainable Development of Intangible Cultural Heritage with AI: Cantonese Opera Singing Genre Classification Based on CoGCNet Model in China. *Sustainability* **2022**, *14*, 2923. <https://doi.org/10.3390/su14052923>

Academic Editor: Asterios Bakolas

Received: 8 February 2022

Accepted: 27 February 2022

Published: 2 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Cantonese opera, which is a representative genre of drama in Guangdong Province in China, was selected as one of the first representative lists of China's national intangible cultural heritage in 2006 and was included in the UNESCO Representative List of the Intangible Cultural Heritage of Humanity in 2009. As a style of drama sung in Cantonese, Cantonese opera has a history of over 300 years, showing both deep traditional cultural heritage and strong Chinese identity. Its artistic, cultural, social and economic values are culturally and viscerally supportive to the development of cultural and creative industries. The sustainable development of Cantonese opera can not only lead to a more scientific and standardized heritage and innovation of Cantonese opera culture, but also promote the flourishing of cultural and creative industries. Taking advantage of artificial intelligence technology to rescue, excavate, organize, protect and disseminate traditional culture has become the direction of intangible cultural heritage conservation today [1,2]. Besides, studying the intelligent development of intangible cultural heritage will also become an

essential issue in the heritage of contemporary excellent traditional culture in China [3,4]. In the art of Cantonese opera, singing music is an important means of shaping characters and expressing the happiness, anger and sorrow of the characters. Hence, establishing a database of Cantonese opera materials, analyzing the characteristics of the singing genres and classifying them by means of intelligent technology are of great significance for the inheritance and development of Cantonese opera culture [5]. In this study, AI technology is combined with Chinese Cantonese opera singing genre classification, so as to change the complicated, diverse and scattered situation of traditional Cantonese opera singing genre classification. Traditionally, Cantonese opera data has been classified and managed manually, with experts relying on personal experience and perceptual factors to classify and identify singing genres. However, this method is not only professionally demanding, but also difficult to guarantee classification results. While, in this study, singing styles are systematically and comprehensively analyzed based on science and technology, then accurate classification and identification on these types are conducted, which results in highly accurate classification results supported by a large amount of data. By using deep learning methods to achieve the intelligent classification of Cantonese opera singing genres, the inefficiency, high consumption, high error rate and other problems of traditional management methods can be solved to a certain extent. In addition, the research is not only beneficial to the scientific management and protection of Cantonese opera data, but also conducive to the improvement of the basic theory of Chinese traditional music and the reform of Cantonese opera teaching, so that the intangible cultural heritage culture can be inherited and innovated in the information era.

As one of the regional operas, Cantonese opera shows different singing styles in each genre [6,7]. And these singing genres have certain similarities due to their commonality in language, musical instrument and culture. Hence, this study attempts to identify Cantonese opera singing genres by using audio classification techniques after conducting the feature analysis of Cantonese opera genres. Although relatively few studies on the analysis of Cantonese opera singing cadences have been carried out, more studies on music genre/audio classification have been conducted. To be specific, the article [8] used an integrated extraction of features and classifiers to recognize dialects; Wu [9] proposed a task-independent model FreqCNN; Cao Yi [10] put forward a dual-feature 2nd order dense CNN noise-robust urban audio classification model. Besides, some studies have shown that the extraction and selection of features have an impact on model performance. In detail, Ye [11] utilized CycleGAN to conduct style migration on the joint features of CQT features and Mel spectrum; The article [12] employed parallel input training models to improve classification performance; Gajanan [13] performed speech/music classification based on IIR-CQT sound spectrograms. The researchers extracted the frequency-domain features of audio signals and made use of CNN to learn the depth features of each signal so as to achieve the classification [14–16]. The article [17] presented a combination of one-dimensional convolutional and bidirectional recurrent neural networks for music style classification; The article [18] used CNN as a feature extractor. Considering the logical correlation on time series, Sylvain [19] brought up a method based on sequence pattern mining technique to extract relevant features; Wang [20], Jia [21] and Xia [22] applied LSTM to their respective classification techniques. Furthermore, other researchers have applied deep learning-based audio classification techniques to environment, life and emotion classification. Zhan [23] raised an accent recognition method on the basis of hybrid speech features; Mishachandar [24] proposed deep neural network architecture for recognizing cetaceans, fish, marine invertebrates, man-made sounds and other natural marine noises; Lhoest [25] came up with a model for ambient sound recognition; The article [26] made advantage of DCNN to extract speech emotion features and then classify them for recognizing of the emotional state of the speakers; The article [27] provided a detailed review of music information retrieval techniques. The classification of Cantonese opera, which belongs to the combination of music and speech, can be based on deep learning-based music/audio processing methods. Huang [28] proposed a time-frequency feature extraction method

based on variable Q-transformation and filter bank techniques with a multilayer cascaded neural network for opera classification. This indicates that it is feasible to use deep learning to classify Cantonese operas. Still, the unique musical characteristics of Cantonese opera genre singing and the existence of contextual dependencies between its fragments have not been analyzed and studied yet, so utilizing a deep learning-based approach for Cantonese opera genre singing analysis is still full of challenges.

Based on the above research, a Cantonese opera singing genre classification method on the basis of CoGCNet model is proposed. This method uses a hybrid structure based on multilayer cascaded CNN and LSTM to fuse the shallow and deep features of each Cantonese opera singing genre, meanwhile taking the contextual relevance among Cantonese opera singing genres into account [29,30]. The experimental results on the constructed dedicated datasets CoD1 and CoD2 show that the method is effective in the classification of Cantonese opera singing genres. The remainders of this paper are structured as follows: In Section 2, the research data analysis and collection are described; In Section 3, the knowledge related to Cantonese opera and the theory of spectral analysis, etc. are introduced; In Section 4, the experimental setup related to this study is described; In Section 5, the experimental results are analyzed; And the conclusions and outlook of the study are given in Section 6.

2. Database

2.1. Data Analysis

A dedicated database, which is collected for the main singing genres in Cantonese opera, is presented in this study, which is collected for the main singing genres in Cantonese opera. This work fills the gap in the data set for the classification of Cantonese opera singing genres, which is briefly described below.

Cantonese opera is a style of drama sung in Cantonese, the representative element of which is the singing voice. This opera, incorporates a variety of musical and dramatic elements and combines the sounds of bangzi (a kind of percussion instrument) and erhuang (Chinese opera singing) with the Cantonese dialect. It has creatively expanded the artistic expression of Chinese opera, becoming a masterpiece of the art of the northern and southern Chinese opera, very different from other Chinese opera genres. In terms of the development, Cantonese opera had the same or similar singing style, and was sung by different actors and actresses. While, distinctive personalities and styles were produced due to the different vocal conditions and singing techniques and habits of these actors and actresses. Finally, it was developed into a more fixed form and was accepted and loved by the audiences, thus forming a genre of singing. In various periods of history, different singing styles have emerged, of which “Xia qiang” “Hong qiang” “Ma qiang” “Xinma qiang” and “Fan qiang” are all representatives of the current genres. Specifically, the “Xia” shows a thick and sweet sound, especially in the middle and bass regions, with a strong resonance, and the lines of this singing style are unadorned but grand and flamboyant; The “Hong” is unique in its sweetness, crispness, roundness, moistness, delicacy and wateriness; The “Ma” melody is jumpy, half-singing and half-white, with short, powerful, lively and comical lines and a mixture of dialects and colloquialisms; The “Fan” melody is bright and rounded, with a clear rhythm and a lot of skipping, without deliberately pursuing the coherence of the singing, and good at using a distinctive rhythmic drag, etc. In order to visualize the differences among these various schools of singing, the Mel spectrum of the five types of the typical singing voice is shown in Figure 1.

Figure 1. Mel-Frequency Cepstrums of each genre.

2.2. Data Set

The data used for the experiments in this paper is a dedicated dataset, with different Cantonese opera singing genres as the labels of the dataset, and the type of data labels is determined by the popularity and diversity of Cantonese opera singing genres. In this paper, various factors are integrated to collect ten types of singing genre represented by Luo Jiabao, Hong Xiannv, Ma Shizeng and others, which were labelled as Luo, Hong, Ma, Deng, He, Bai, Chen, Lpc, Xue, Gui respectively. The data were sourced from the internet platforms, including China Opera Network. The dataset collected audio from different scenes, environments and times for each Cantonese opera singing genre. The data contained as many typical Cantonese opera singing voices as possible, and was used to remove the impact of the audio sound quality in different periods on the classification effect. In most cases, a complete Cantonese opera performance requires the collaboration of several actors, thus continuous singing passages from each singing genre need to be collected as valid experimental data. From a large amount of Cantonese opera audio collected, multiple Cantonese opera pieces with varying lengths but same total duration were selected for each singing genre so as to ensure the balance of the data. The experimental data consisted of 305 Cantonese operas with a total of 1000 cantos in WAV format, each with a length of 30 s and a sampling rate of 44,100 Hz. The relevant data information is shown in Table 1.

Table 1. Information of data label.

Num	Class	Label	Operas Num	Number of Files (.Wav)
1	Luo Jiabao	Luo	34	100
2	Hong Xiannv	Hong	32	100
3	Ma Shizeng	Ma	34	100
4	Deng Zhiju	Deng	29	100
5	He Feifan	He	33	100
6	Bai Jurong	Bai	27	100
7	Chen Xiaofeng	Chen	31	100
8	Luo Pinchao	Lpc	28	100
9	Xue Juexian	Xue	29	100
10	Gui Mingyang	Gui	28	100

The model is divided into CoD1 and CoD2 based on the similarity of each singing genre and the quality of the audio signal. In detail, CoD1 contains five typical singing genres, all of which are the original Cantonese opera singing signals. And CoD2 includes ten Cantonese opera singing genres that have been manually processed and data-enhanced.

With manually processed, the CoD2 reduces the interference of noise and selects more representative Cantonese opera audio for classification. The information of datasets is shown in Table 2. Experiments were conducted on both datasets so as to verify the classification performance of the model.

Table 2. Information of datasets.

Dataset	Labels	Num	Train/Test
CoD1	Luo, Hong, Ma, Deng, He	165	70%/30%
CoD2	Luo, Hong, Ma, Deng, He, Bai, Chen, Lpc, Xue, Gui	305	70%/30%

3. Methods

With regard to classification tasks, feature extraction and classifier learning are two essential processes. The raw data is often large and mixed, and can't be used directly for classification due to its inapparent features, so salient features that can represent the sample data need to be extracted. Audio files contain plenty of time-domain and frequency-domain features, and before training the model with a neural network, feature extraction is first performed. Considering the different styles of Cantonese opera singing genres, the Mel frequency spectrum of Cantonese opera signals is used as the feature vector in this study, and feature extraction is integrated into and through the entire training process of the classification model through deep learning methods, so as to realize the analysis on Cantonese opera singing features and the classification of singing genres. The network structure of the classification model involved is mainly the combination of CNN and LSTM, using a multilayer cascaded network architecture. The model fuses the shallow and deep information of each singing section by means of the cascaded fusion of the first-level network (Inception-CNN), so as to enhance the model's capacity to extract the singing genre features; The second-level network, which is a combination of CNN and two-layer LSTM stacking, learns the logical features between each singing section and extracts contextual association semantics. LSTM, which introduces a sigmoid function based on RNN, effectively handles the short-term dependency problem while solving the gradient disappearance problem with RNNs in general through the cell state memory information. Compared with other network structures, the model combining CNN and LSTM proposed in this study has better comprehensive performance in the classification of Cantonese opera singing genres.

3.1. Cantonese Opera

In addition to the unique vocal system, Cantonese opera also shares the attributes that are common to other forms of music, such as pitch, intensity, timbre and rhythm [31]. To be specific, pitch is determined by the frequency of vibration of the vocal body, indicating the level of pitch. Besides, different pitches naturally convey different emotions. For instance, high-frequency tones express positive emotions, such as urgency, brightness and excitement. While, low-frequency tones convey negative emotions, such as laziness, depression and decadence, etc. In addition, the range of pitches varies from instrument to instrument. Intensity refers to the loudness of a sound, which is determined by the amplitude of the vibration of the sound source and is mainly used to distinguish the loudness of a sound. It is commonly measured in “decibels” and expressed in terms of energy magnitude. Timbre, which means the characteristic properties of a sound, is determined by the structure of the material and the mechanism of sound production, and is utilized to distinguish among different types of the sound. It naturally varies according to the instrument used. The rhythm, which is based on the beat, is the variation in strength and weakness of the music. And it is the difference in the rhythm, beat and rhythmic strength and weakness of the rhythm that creates different musical styles [11].

3.2. Spectrum Analysis

When studying the audio/music style classification, the corresponding audio files are usually used as input data. In consideration of the inconsistency in size and duration of the original audio files, the audio signal is first analyzed, including slicing, framing, windowing, time-frequency transformation of the signal (Fourier transform) and the extraction of time-frequency features (Mel-Frequency Cepstrum).

3.2.1. Slicing

Slicing means the cutting of an audio file into multiple segments with equal length. Through slicing, the features corresponding to each frame are processed to obtain fragment-level features, thus increasing the number of input samples. To a certain extent, the data is augmented, which is conducive to improving the generalization capacity of the model.

3.2.2. Framing

As non-smooth signals, audio signals are produced by the vibration of the vocal organs, but a number of uncontrollable factors usually influence the state of motion of the organs. Scientific research has shown that the vibrational audio of the muscles of the human vocal tract is limited, and the spectral characteristics of the sound signal tend to be nearly smooth over very short period of time. In view of the continuity within the Cantonese opera singing signal, overlapping framing is generally adopted so as to avoid a certain loss of information caused by physical isolation. As shown in Figure 2, the framing process of a 30-s Cantonese opera audio sample is illustrated with a duration of 1.5 s and an overlap rate of 50%, i.e., 50% of the overlap between every two adjacent frames, discarding the last part of the sample that is less than the duration.

Figure 2. Schematic diagram of Cantonese opera signal framing.

3.2.3. Windowing

Windowing is essentially a numerical operation on the segment obtained through framing, which means that the signal is smoothed to reduce the error caused by framing, making the overall situation more continuous and avoiding the Gibbs effect [32,33]. The commonly used window functions include the Rectangular window and the Hamming window.

The formula for a Rectangular window is as follows:

$$\omega(n) = \begin{cases} 1 & 0 \leq n \leq L-1 \\ 0 & \text{others} \end{cases} \quad (1)$$

The formula for a Hamming window is described below:

$$\omega(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{L-1}\right) & 0 \leq n \leq L-1 \\ 0 & \text{others} \end{cases} \quad (2)$$

The above two windows exert different influence on the signal spectrum. On the one hand, the Rectangular window is also a zero power window from the time variable de-analysis. There is a certain degree of risks of causing high-frequency interference and spectral leakage due to the sudden truncation of the signal at both ends of the transformation process. On the other hand, the Hamming window increases the width of the main flap, decreases the height and has a smaller side flap, which effectively suppresses the

spectral leakage while maintaining the high-frequency characteristics of the signal. This study uses the Hamming window on the basis of overlapping framing. By adding the window, the original audio signal, which lacks periodicity, takes on some characteristics of a periodic function, thus making preparation for the next step of Fourier expansion.

3.2.4. Fast Fourier Transformation

Fast Fourier Transformation (FFT) converts a coefficient representation into a point-value representation by multiplying it by a coefficient representation. To be specific, a signal is first transformed into the frequency-domain by means of a discrete FFT, and then an inverse discrete Fourier transform is performed to synthesize the original signal, i.e., the process of first finding the value and then interpolating it. This algorithm uses a partitioning idea, partitioning to find the value of the entire polynomial when $x = \omega(k)[n]$, then transforming the original $O(n^2)$ complexity of the plain polynomial multiplication into an $O(n \log n)$ algorithm.

3.2.5. Mel-Frequency Cepstrum

Audio signals are originally one-dimensional time-domain signals, and it is intuitively difficult to see the frequency change pattern of such signals. While, compared with the time-domain signal directly sampled from the audio signal, the frequency-domain signal, where the frequency is transformed into a vector after some variation of the audio signal, is closer to the hearing mechanism of the human ear [34]. Thus, the original audio signal can be transferred from the time-domain into the frequency-domain, and then the feature extraction can be conducted on relevant signals. Mel-Frequency Cepstrum, which was proposed based on Mel frequency, is more in line with the hearing system of the human ear. It belongs to the frequency-domain features, is mainly used for audio data feature extraction and reduces the dimension of operation. The correspondence between Mel frequency and Hz frequency is shown in Equation (3). And the frequency-domain features of relevant signal can be calculated through the mapping relationship between the two frequencies.

$$mel = 2595 \log_{10}(1 + f/700) \quad (3)$$

3.3. LSTM

Long Short-Term Memory (LSTM), which is a kind of temporal recurrent neural network, takes the temporal characteristics of the input vector into account to extract temporal contextual features [35]. As a branch of RNN, LSTM adds memory blocks inside each neuron in terms of structure compared with RNN, i.e., input gate, output gate and forgetting gate. LSTM not only inherits the advantages of RNN in mapping input and output considering contextual information, but also solves some problems, such as “long-range dependence” and “gradient disappearance” when processing long sequences. These problems are the typical difficulties for RNN faces in remembering information at long intervals when computing implicit temporal layers. The structure of LSTM is shown in the following equations from (4) to (9):

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (4)$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (5)$$

$$\tilde{C}_t = \tanh(W_{\tilde{C}}[h_{t-1}, x_t] + b_{\tilde{C}}) \quad (6)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (7)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (8)$$

$$h_t = o_t * \tanh(C_t) \quad (9)$$

where W_f , W_i , $W_{\tilde{C}}$ and W_o refer to the weight parameters, b_f , b_i , $b_{\tilde{C}}$ and b_o mean the biases, and x_t is used as the input sequence that, in combination with the state of the hidden layer

h_{t-1} above, forms the forget gate f_t via the activation function. Besides, input gate i_t and output gate o_t are also calculated from h_{t-1} and x_t . And the forget gate f_t is united with the pre-unit state C_{t-1} to determine whether to discard the message.

4. Experimental Setup

The overall framework of the CoGCNet model based Cantonese opera singing genre classification method proposed in this paper is shown in Figure 3, which is divided into the following two main stages: Cantonese opera singing data processing and model design and prediction.

- (1) Data processing. It consists of data pre-processing and data augmentation. After a series of operations, such as framing, windowing and Fourier transform of the original Cantonese opera singing signal, the Mel-Frequency Cepstrum of each Cantonese opera singing signal was then obtained as model input by a Mel filter.
- (2) Model design and prediction. The CoGCNet model based Cantonese Opera singing genre classification model consists of a first-level network (Inception-CNN) and a second-level network (CNN-2LSTM). This classification model takes advantage of the cascaded fused Inception-CNN network to fuse the shallow and deep information of each singing section so as to enhance its feature extraction capacity. The CNN-2LSTM network, which is a combination of CNN and two-layer LSTM stacking, learns the logical features among each audio segment to extract the contextual association semantics. Finally, the Majority Vote Algorithm is utilized to decide the predicted Cantonese opera singing category and output the results.

Figure 3. Flow chart of model processing.

4.1. Data Processing

The original Cantonese opera repertoire was sliced and transformed into a number of Cantonese opera segments with equal duration (30 s) as the model sample set, and was randomly divided into 70%/30% as the training set and test set. The Cantonese opera audio signals were divided into short-time signals by framing, windowing, and mapping the audio signals from the time-domain to the frequency-domain through conducting the Fourier transform on each frame to obtain the spectrogram. Finally, the results of each frame were stacked, i.e., the corresponding spectrum along another dimension, to obtain the time-varying spectrogram. The Mel-Frequency Cepstrum was then obtained by a Meier filter, which involves the sampling parameters shown in Table 3.

Table 3. The sampling parameters of Mel-Frequency Cepstrum.

Num	Name	Parameter	Value
1	sampling_rate	sampling rate	44,100 Hz
2	duration	duration	30 s
3	n_mels	Number of Mel filter banks	128
4	n_fft	FFT Window length	1024
5	hop_length	Window sliding length per frame	128
6	spe_width	The duration of the spectrum map	1.5 s

In order to demonstrate the feature engineering more intuitively, the data processing process is visualized below, as shown in Figure 4.

Figure 4. The process of obtaining the Mel-Frequency Cepstrum.

4.2. Network Architectures

In this section, the CoGCNet model neural network was used to further extract the higher-level features of the spectrogram. And a 70% sample set was employed as the training set to train the network model which includes convolutional layers, ReLU layers, Max-Pooling layers, Concat layers, LSTM layers, Flatten layers, Dense layers, Dropout layers, Batch Normalization layers, etc. The CoGCNet model is composed of two-level neural networks, as shown in Figure 5 and Table 4. The Mel-Frequency Cepstrum corresponding to the Cantonese opera singing segments was applied as the model input. The shallow and deep fusion feature expressions of each Cantonese opera singing segment were obtained through a first-level network that consists of CRMD block and MsFF_block structures, based on the Cantonese opera repertoire singing segments within the shallow and deep layers. The second-level network consist of a CNN and a stacked LSTM, which is utilized to learn the semantics of contextual associations among individual singing segments and the importance of logical relationships among individual singing segments within the same Cantonese opera repertoire.

Figure 5. The network structure of CoGCNet model.

Table 4. Hyperparameters of the CoGCNet model.

Block	Hyperparameters
CRMD	Conv2D (kernel size = (3, 3), strides = (1, 1)); MaxPooling2D (pool_size = (2, 2), strides = (2, 2)); Dropout (discard rate = 0.25)
CBR	Conv2D (kernel size = (3, 3), strides = (2, 2))
LSTM	LSTM1 (units = 256, return_sequences = True); LSTM2 (units = 128, return_sequences = False)
Dense	Dense1 (units = 512, activation = 'relu', kernel_regularizer = tf.keras.regularizers.l2 (0.02)); Dense2 (units = 10, activation = 'softmax', kernel_regularizer = tf.keras.regularizers.l2 (0.02))

4.2.1. First-Level Network Model Design

The first-level network consists of two modules, the two-dimensional CRMD block and the MsFF_block (Multi-scale Feature Fusion block), which learn the frequency-domain characteristics within each Cantonese opera chant and fuse the deep and shallow features of each chant. The design of the CRMD block and MsFF_block is described below.

1. CRMD block

In CNNs, the convolution methods include one-dimensional convolution and two-dimensional convolution. Compared with time-domain signals, frequency-domain signals are closer to the hearing mechanism of the human ear. In this paper, a CRMD block is proposed based on a two-dimensional convolution so as to transfer the audio signals from the time-domain to the frequency-domain. Then, the two-dimensional convolution is utilized to extract higher-level features from both time and frequency directions, respectively. The CRMD block, which is the basic component of the CoGCNet model network structure, is composed of a Conv (convolution) with a stride of 1×1 and a convolution kernel size of 3×3 , a ReLU activation function, Max-Pooling and Dropout.

With regard to general classification models, the Dropout layer is usually added to the fully-connected layer so as to prevent overfitting and improve the generalization capacity of the model. As the convolutional layer has fewer parameters, so the addition of Dropout layer exerts little effect. The lower layers are noisy, while the higher layers, which are fully connected layers, can increase the robustness of the model and generalization performance.

When designing the network, each layer of neurons is set to represent a learning feature, i.e., a combination of several weights, with all neurons in the network acting together to characterize specific attributes of the input data. In consideration of the small and disorganized resources of the Cantonese opera repertoire, the effective data set that can be collected and collated is on the small side, the amount of data is too small relative to the complexity of the network. i.e., the expressiveness and fitting capacity of the network, and overfitting occurs, i.e., there are many repetitions and redundancies among the features represented by each neuron. The CRMD block structure introduced in this section, with the Dropout layer following the convolutional layer, avoids the overfitting problem of the network model to some extent.

2. MsFF_block

It is difficult to accurately distinguish the categories of Cantonese opera singing voices through a single frequency-domain feature alone, due to the similarity of the musical characteristics of different Cantonese opera singing voices. Hence, the MsFF_block structure is proposed in this paper on the basis of a lightweight Inception structure [36] combined with CBR block composition. As the basic component of the MsFF_block structure, the CBR block contains Conv (Convolution), Batch Normalization (BN) and the ReLU activation function. In terms of the overall module structure a symmetric two-layer stacked convolutional kernel structure is adopted to split the larger convolutions and extract features from different layers for fusion so as to achieve a multi-dimensional feature representation. Finally, a Concat layer is employed to connect the features in each dimension, which maintains a certain amount of feature space information while enhancing the generalization capacity of the model to the feature range.

4.2.2. Second-Level Network Model Design

Although typical CNNs can effectively learn frequency-domain features within Cantonese opera singing, they have difficulty in extracting temporal contextual features. Thus, in this paper, a second-level network is put forward, which is a feature cascade stacking network consisting of a double-layer LSTM and convolutional layers. By increasing the depth of the network, the time to extract features can be extended. However, the stacking of multiple LSTM layers will lead to the problem of “gradient disappearance” due to the inability to iteratively update the shallow weights. Therefore, this level of network adopts a stacked network of double-layer LSTM to extract deep separability features layer by layer, from the basic conceptual features to more abstract deep features. Besides, the stacked convolutional kernel extends the range of the spectral features learned in the singing section, allowing the network to have an enhanced capacity to recognize potential patterns in the acoustic spectrum. Each layer outputs a sequence of feature vectors as an input to subsequent layers, which are chunked over time to observe and represent features at different time scales, thereby enhancing the representational power of the model.

4.3. Training Algorithm Settings

The CoGCNet model used the first-level network to obtain the fused feature expressions of the shallow and deep layers within each Cantonese opera chant, and then utilized the second-level network to acquire the contextual association semantics among the chants based on time series relationships. Finally, the majority voting algorithm was employed to predict the category labels for Cantonese opera singing. The Adam algorithm was used for network optimization, with a learning rate size of 0.00001 and a batch size of 128. The hardware and software environment related to this experiment is as follows: the processor is Core i7-9700@3.00GHz octa-core CPU, 16.0 GB RAM, the graphics card is NVIDIA GeForce RTX 2060 (6 GB) GPU; The operating system is 64-bit window10, python version 3.7.10, deep learning framework TensorFlow-GPU 2.2.0, Keras 2.6.0, music processing framework librosa 0.8.1, etc., and CUDA version 10.1 parallel computing framework with CUDNN version 7.6.5 deep neural network acceleration library is used.

4.4. Metrics

In this study, the ablation experiments and comparison experiments were conducted for the classification model of the Cantonese opera singing genre, and the evaluation metrics were calculated from the confusion matrix of each experiment. The classification model proposed in this study was then validated and evaluated with several different metrics. Specifically, regarding the classification model and related experiments, the feasibility was assessed by using the precision value P (Precision), the recall rate R (Recall) and the $F1$ value ($F1$ measure) shown in the following equations from (10) to (12).

$$P = \frac{TP}{TP + FP} \times 100\% \quad (10)$$

$$R = \frac{TP}{TP + FN} \times 100\% \quad (11)$$

$$F1 = \frac{2PR}{P + R} \times 100\% \quad (12)$$

where P refers to the accuracy rate, R represents the recall rate; TP means the number of true positive samples, FP denotes the number of false-positive samples, and FN shows the number of false sub-samples.

5. Experimental Results

In this section, the ablation experiments of the CoGCNet model, the comparison experiments with existing techniques and the related experimental results are described and analyzed, and then the effectiveness of the CoGCNet model in classifying Cantonese opera cantatas is validated.

5.1. Ablation Experiment

The two core modules of the CoGCNet model were replaced by the ordinary convolutional network structures in order to analyze the impact of the two core building blocks, namely the MsFF_block structure in the first-level network and the double-layer stacked LSTM structure in the second-level network, on the classification performance of the model. And the specific comparative structure of the two core modules of the CoGCNet model on the dataset CoD1 is shown in Table 5, the results of the ablation experiment are shown in Table 6, and a comparison of the Accuracy value variation curves is shown in Figure 6. Each model was trained on COD1 for 150 epochs. To be specific, the learning rate was set to 0.00001, the batch size was set to 128, the ratio of the training set to the validation set was 7:3, the categorical cross-entropy loss function was used as the model training performance indicator, and the adaptive moment estimation (Adam) optimizer was utilized for parameter adjustment and optimization during the training process.

Table 5. Comparative structure of the ablation experiment.

Number	CNN	2LSTM	MsFF	Model
A	✓	✓	—	CNN+2LSTM
B	✓	—	✓	CNN+MsFF
C	✓	✓	✓	CoGCNet

Table 6. Classification results of each network structure.

Model	F1/%					F1/%	P/%	R/%
	Luo	Hong	Deng	He	Ma			
CNN+2LSTM	82.64	79.04	78.26	96.08	99.01	87.01	89.38	87.20
CNN+MsFF	86.81	86.89	86.15	93.43	99.00	90.46	90.80	91.00
CoGCNet	87.56	92.46	90.25	96.55	99.01	93.16	93.18	93.20

Figure 6. Accuracy curve of each model.

The categorical cross-entropy loss is calculated as follows:

$$loss = - \sum_{i=1}^n \hat{y}_{i1} \log y_{i1} + \hat{y}_{i2} \log y_{i2} + \dots + \hat{y}_{im} \log y_{im} \quad (13)$$

where y represents the true value, $\hat{y}$ represents the predicted value, n means the number of samples, and m denotes the number of data categories. When the predicted value is equal to the true value, the loss function value is 0. The greater the difference between the two is, the greater the loss value is.

Table 6 shows that the combined use of the two core modules, is conducive to improving the model performance. According to the training test results, although there was a slight decrease in the $F1$ value of model B (CNN+MsFF) compared with that of model A (CNN+2LSTM) for He and Ma, the overall $F1$ value, Precision and Recall rate reached 90.46%, 90.8% and 91.00%, exceeding 3.45%, 1.42% and 3.8%, respectively, with more obvious advantages in $F1$ value and precision. Thus, the model with only the MsFF_block structure was proved to be more effective in classifying Cantonese opera singing genres, and the features of Luo, Hong and Deng could be extracted in a deeper and more multi-dimensional way. In contrast, the model with only a double-layer stacked LSTM structure showed less obvious performance advantages and was ineffective in extracting contextual association features. While, model C (CoGCNet model), which used both MsFF_block structure and double-layer stacked LSTM structure, achieved 87.56%, 92.46%, 90.25% and 96.55% for Luo, Hong, Deng and He respectively, while maintaining the high $F1$ value of 99.01% for Ma compared with model A and model B. Besides, the overall $F1$ value and accuracy of the model are also high. Meanwhile, the overall $F1$ value, Precision and Recall of model C reached 93.16%, 93.18% and 93.20%, all within a range of 2.2% to 6.16% improvement over the previous two models, which demonstrates that model C has the capacity to cascade and fuse deep, multi-dimensional features and contextual association features. There is a significant enhancement effect on the classification results of the model.

Figure 6 indicates that the accuracy of model C has been in the range between model A and model B from the beginning of the training. After 30 iterations, the accuracy of model C starts to show a tendency to surpass that of the two models. Then after a few dozen iterations, the accuracy of model C has stabilized and continues to maintain the optimal accuracy. It is proved that model C is optimal for the classification of Cantonese opera

singing genres. This shows that using only the MsFF_block structure in the primary network or the double-layer stacked LSTM structure in the secondary network has some limitations on the performance improvement of the model while the other network structures remain consistent. In other words, there is a logical relationship between the structure of the multi-dimensional feature fusion in the primary network and the contextual association features in the secondary network, which in turn validates the effectiveness of the design of the CoGCNet model structure on the Cantonese opera singing classification task.

5.2. Performance of CoGCNet in 10-Class Dataset

In order to analyze the training performance of the CoGCNet model on different datasets, 150 Epoch training sessions were conducted on the ten-class Cantonese opera singing dataset CoD2 in this paper. To be specific, the learning rate was set to 0.00001, the batch size was set to 128, the ratio of the training set to the validation set was 7:3, the training process had a reverse gradient, and the validation process had no reverse gradient. With regard to the ten-category multi-classification problem, after the feature extraction in the convolutional layer, the softmax function needed to be connected to output the distribution probability, so the categorical cross-entropy loss function was also used as the model training performance indicator, and the adaptive moment estimation (Adam) optimizer was utilized for parameter adjustment and optimization during the training process.

As shown in Figure 7, the loss values of the validation set and training set of the CoGCNet model fit fast in the first 20 iterations, rapidly became smaller, and then slowly decreased. The loss value of validation set fluctuated more than that of the training set, which was smaller but converge at a similar rate, and both had an excellent convergence effect. Finally, the loss value of training set stabilized between 0 and 1, with only slight oscillations, indicating the end of the model fitting. The accuracy of the training set and validation set of the CoGCNet model grew faster in the first 40 epochs and then grew slowly. The final accuracy of validation set was between 94–95%, and that of training set was between 96–97%, which showed that the training set had a higher accuracy. Still, the overall growth trend of both sets was consistent during the training process, and the absolute accuracy was stable, which proves that the model training achieved better results.

Figure 7. Iteration curves of Accuracy and Loss.

To verify the classification performance of CoGCNet model on the ten-class dataset CoD2, five training sessions were conducted by applying the same CoD2 in the same environment. The dataset was randomly divided into 70%/30% as the training set and test set each time. The Precision, Recall and F1 value of each training session were recorded, and the mean values of the five cross-validations were collated.

As can be seen from Table 7, the CoGCNet model achieved a better classification on CoD2, with Precision of 95.68%, Recall of 95.58% and F1 value of 95.60%. He sings used clever rhythmic dragging to create a very personal singing genre, without the deliberate pursuit of vocal continuity. He achieved the best predictions of 99.95%, 98.02% and 99.00% for Precision, Recall and F1 value, respectively. The dragging voice of Ma is self-contained, with vocal rhythms that are more guttural than guttural. Besides, it also has highly recognizable vocal characteristics, so its prediction was better than that of other classifications, with Precision and F1 value of 99.95% and 98.46%, respectively. Although Luo has its own singing style, it also draws on the strengths and characteristics of other singing styles. Hence, its prediction was slightly lower than that of other classifications, with the Precision and F1 value of 86.79% and 89.76%, respectively. Hong had the lowest Recall value of 91.92%, but its Precision reached 95.79%, indicating that the model had more missed detections for Hong's classification. In addition, the Precision of CoD2 is higher than that of CoD1, which proves that our data enhancement has improved the quality and is more in line with the requirements of Cantonese opera classification. In summary, the CoGCNet model performed well on the ten-class Cantonese opera dataset CoD2, and showed a strong feature extraction capacity to meet the requirements of Cantonese opera cantata classification.

Table 7. Performance metrics of CoGCNet model in CoD2.

Num	Class	Precision (%)	Recall (%)	F1 (%)
1	Luo	95.19 ± 0.48	99.00 ± 0.50	97.06 ± 0.48
2	Lpc	93.40 ± 0.88	99.00 ± 0.01	96.12 ± 0.99
3	Hong	95.79 ± 2.06	91.92 ± 3.54	93.82 ± 1.64
4	Gui	99.93 ± 0.04	94.95 ± 1.01	97.38 ± 1.76
5	Ma	99.95 ± 0.03	97.03 ± 0.49	98.46 ± 0.98
6	Deng	93.88 ± 2.01	92.00 ± 0.97	92.93 ± 1.94
7	Xue	97.96 ± 1.02	96.00 ± 1.00	96.97 ± 1.50
8	Luo	86.79 ± 5.47	92.93 ± 3.03	89.76 ± 0.48
9	He	98.02 ± 0.49	99.95 ± 0.48	99.00 ± 0.48
10	Chen	95.88 ± 0.80	93.00 ± 3.50	94.42 ± 2.03
-	Average	95.68 ± 0.76	95.58 ± 0.75	95.60 ± 0.74

In order to visualize the classification effect of the model under each label, the Confusion matrix was used to visually judge the classification effect obtained by the model on the test set for Cantonese opera singing. As shown in Figure 8, the predictions of Bai, Lpc, He and Ma were nearly accurate. They achieved very satisfactory classification results, while the classification results of the rest of the singing labels had a few cases of misclassification. For example, Chen was easily predicted as Luo due to the similarity in the accompaniment and musical instruments between the two. In general, the CoGCNet model proposed in this paper meets both the design requirements and the usage needs on classifying Cantonese opera cadences, and the performance is promising.

Figure 8. Confusion matrix of Cantonese opera singing classification.

5.3. Discussion and Comparative Analysis

In the previous section, the effectiveness of the two core modules of the CoGCNet model in classifying Cantonese opera singing genres was demonstrated through ablation experiments. In this section, the classification performance and computational cost will be combined to compare the CoGCNet model and other commonly used deep learning models. These models will be trained with the same data pre-processing, and when the learning rate is 0.00001 and epochs are 150, the models tend to be stable. And the optimal value will be adopted as the result of the comparison experiments after multiple experiments. The comparison models include AlexNet, VGG16, VGG19, GoogLeNet, ResNet50 and GoGCNet model, and the experimental results are shown in Table 8 below.

Table 8. Classification results of each network structure.

Model	Recall (%)	Precision (%)	F1 (%)	Parameters (M)	FLOPs (G)	CPU Test Time (ms)	GPU Test Time (ms)
AlexNet	89.52	89.6	89.43	162.46	1.35	47.98	26.88
VGG16	91.98	92.51	92.00	408.32	20.60	118.51	62.06
VGG19	92.04	92.51	92.50	428.59	26.20	128.11	65.83
GoogLeNet	92.75	93.30	92.85	22.78	1.48	13.14	7.63
ResNet50	95.00	95.07	94.98	122.41	5.19	37.48	22.16
CoGCNet	95.58	95.69	95.60	37.56	0.40	7.44	4.05

As shown in Table 8, the CoGCNet model achieved Recall of 95.58%, Precision of 95.69% and F1 value of 95.60% on dataset CoD2. Compared with other commonly used CNN models, the CoGCNet model had the improvement of 0.58~6.06%, 0.62~6.09% and 0.62~6.17%, respectively, showing the best classification evaluation indexes for all classification results. By comparing the number of parameters and floating-point operations per second (FLOPs) of each model, it can be seen that the CoGCNet model had 37.56 M parameters and 0.4 G FLOPs and had the lowest number of parameters and complexity. The number of FLOPs was reduced by 25.8 G, and the number of parameters was reduced

by 391.03 M, which was 98.47% and 91.24%, respectively. Besides, the CoGCNet model has the lowest FLOPs and complexity than the smallest GoogLeNet. Although the number of parameters was increased by 14.78 M, the FLOPs were reduced by 1.08 G, the complexity of the model was much smaller than that of GoogLeNet. The CoGCNet model has the best computational performance for the above two indicators. By comparing the model prediction time in different environments, the testing time of the CoGCNet model was 4.05 ms on GPU and 7.44 ms on CPU, both of which were lower than that of other CNN models, with 13.14 ms and 7.63 ms on CPU and GPU, respectively compared with the GoogLeNet model with the smallest number of parameters. Regarding the CoGCNet model, the prediction time was reduced by 43.38% and 46.92%, which minimized the prediction time and achieved the lightweight model. Compared with the ResNet50 model, which has similar model accuracy, the CoGCNet model has about 1/4 of the number of parameters and less than 1/13 of the FLOPs, and the inference speed on both GPU and CPU was reduced by about four times, and all indicators were better than those of ResNet50. It can be seen that the CoGCNet model has better genre feature extraction in Cantonese opera singing spectrograms performance.

Figure 9 shows the relationship between the accuracy and classification prediction time on different neural network models by using GPU on the ten-class dataset CoD2, which can display more intuitively the classification effect and model performance of the six models. The VGG model with two structures is the slowest, but the Accuracy value of this model is better than that of AlexNet. GoogLeNet, as a lightweight model, has a shorter prediction time, but the Accuracy value of this model is slightly inferior to that of ResNet and CoGCNet models. With regard to the ResNet model, which has better classification results, its Accuracy value is better than that of the remaining four types of models but lower than that of the CoGCNet model. In summary, the CoGCNet model achieves the best classification results on the CoD2 dataset, guaranteeing the accuracy of the classification of Cantonese opera singing genres meanwhile improving the prediction time of the model.

Figure 9. Performance of different CNNs.

6. Conclusions

This paper mainly involves the scientific and standardized construction of a database of Cantonese opera audio data and the use of deep learning technique to achieve an intelligent classification of Cantonese opera singing genres. The core of this approach, one of the key tools for achieving the scientific management and sustainable development of in-tangible cultural heritage, lies in the extraction of Cantonese opera singing features and the learning of the intrinsic correlation between the various singing sections. The main task of this study is to make advantage of deep learning techniques to achieve the learning of Cantonese opera singing features and the classification of singing genres. Therefore, a CoGCNet model based classification model for Cantonese opera singing is proposed in this paper. The original Cantonese opera singing signal is pre-processed based on the characteristics of each singing voice so as to obtain the Mel-frequency Cepstrum, which is used as the input of the model. The cascaded CNN is utilized to fuse the shallow and deep features of each singing voice. Then, a hybrid neural network model combining CNN and double-layer LSTM is employed to extract the contextual association semantics. Finally, a majority voting algorithm is used to vote. In terms of model structure, a two-level network is designed in this paper to extract and fuse features within each Cantonese opera singing section and among different singing sections of the same repertoire. This not only extracts the deep and shallow features of the singing sections, but also retains the correlation between the preceding and following sections, thereby making the CoGCNet model more accurate in predicting Cantonese opera singing genres.

The collected experimental data were divided into a 5-class dataset CoD1 and a 10-class dataset CoD2 so as to validate the effectiveness of the classification model on a multi-category problem. The experimental results of the CoGCNet model on datasets CoD1 and CoD2 validated the high performance of the method in both the 5-category and 10-category classification tasks, with both achieving 90% in terms of the *F1* value, Precision and Recall above. In the ablation experiments for the two core modules in the CoGCNet model, the accuracy of the model with the double-stacked LSTM structure was 89.38%, the accuracy of the model with the MsFF_block structure was 90.80%, while the accuracy of the CoGCNet model reached 93.18%. In order to analyze the accuracy of the recognition of each singing genre, the confusion matrix of the classification results of the CoGCNet model on the dataset CoD2 is provided in this paper. In addition, a comparison on the CoGCNet model and the state-of-the-art network models is also conducted in this paper. The accuracy of AlexNet, VGG16, VGG19, GoogLeNet and ResNet50 models were 89.6%, 92.51%, 92.51%, 93.30% and 95.07%, respectively, while the accuracy of the CoGCNet model was 95.69%. In comparison, the CoGCNet model outperformed other algorithms in terms of accuracy, complexity and testing time while minimizing the number of parameters.

The Cantonese opera singing genre classification method proposed in this study is effective on the small sample dataset constructed, meanwhile providing feasible ideas for the research on the sustainable development of other opera cultures. Besides, further optimization of the model on a larger dataset will be considered in the future so as to contribute to the sustainable development of intangible cultural heritage.

Author Contributions: Conceptualization, Q.C. and W.Z.; methodology, W.Z. and Q.W.; software, Q.C.; validation, Q.C., W.Z. and Q.W.; formal analysis, Q.W. and Y.Z.; data curation, Q.C. and Y.Z.; writing—original draft preparation, Q.C.; writing—review and editing, W.Z. and Q.W.; visualization, Y.Z. and Q.C.; supervision, W.Z.; funding acquisition, W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Guangzhou Association for Science & Technology Foundation of China (Grant number. G20210201003); “The First Guangdong-Hong Kong-Macao Greater Bay Area Literary and Artistic Innovation Forum” Key Literary and Artistic Research Topics.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank anonymous reviewers for their criticism and suggestions. We would also like to thank the team lab for equipment support and technical support.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Li, R.Q. Discuss on the Protection and Inheritance of Traditional Performing Arts. *Chin. Cult. Res.* **2019**, 1–14. [\[CrossRef\]](#)
- Song, J.H. Some Thoughts on the Digital Protection of Intangible Cultural Heritage. *Cult. Herit.* **2015**, 2, 1–8.
- Dang, Q.; Luo, Z.; Ouyang, C.; Wang, L.; Xie, M. Intangible Cultural Heritage in China: A Visual Analysis of Research Hotspots, Frontiers, and Trends Using CiteSpace. *Sustainability* **2021**, *13*, 9865. [\[CrossRef\]](#)
- Xia, H.; Chen, T.; Hou, G. Study on Collaboration Intentions and Behaviors of Public Participation in the Inheritance of ICH Based on an Extended Theory of Planned Behavior. *Sustainability* **2020**, *12*, 4349. [\[CrossRef\]](#)
- Xue, Y.F. The Inheritance and development path of Local Opera based on digital resources—Comment on The Digital Protection and Development of Zhejiang Opera Art Resources. *Chin. Educ. J.* **2021**, *10*, 138.
- Liu, J.K. The Inheritance and Development of Cantonese Opera Singing. *Drama Home* **2017**, *10*, 22–24.
- Zhang, Z.L. Cantonese Opera: Connecting the past and innovating the Future. *China Art Daily* **2021**, 2536, 3–4.
- Nbca, B.; Sgk, A. Dialect Identification using Chroma-Spectral Shape Features with Ensemble Technique. *Comput. Speech Lang.* **2021**, *70*, 101230.
- Yu, W.; Hua, M.; Zhang, Y. Audio Classification using Attention-Augmented Convolutional Neural Network. *Knowl. Based Syst.* **2018**, *161*, 90–100.
- Cao, Y.; Huang, Z.L.; Sheng, Y.J.; Liu, C.; Fei, H.B. Noise Robust Urban Audio Classification Based on 2-Order Dense Convolutional Network Using Dual Features. *J. Beijing Univ. Posts Telecommun.* **2021**, *44*, 86–91.
- Ye, H.L.; Zhu, W.N.; Hong, L. Music Style Conversion Method with Voice Based on CQT and Mayer Spectrum. *Comput. Sci.* **2021**, *48*, 326–330.
- Gao, L.; Xu, K.; Wang, H.; Peng, Y. Multi-representation knowledge distillation for audio classification. *Multimed. Tools Appl.* **2020**, *81*, 5089–5112. [\[CrossRef\]](#)
- Birajdar, G.K.; Patil, M.D. Speech and music classification using spectrogram based statistical descriptors and extreme learning machine. *Multimed. Tools Appl.* **2019**, *78*, 15141–15168. [\[CrossRef\]](#)
- Fu, W.; Yang, Y. Sound frequency Classification method based on coiling neural network and Random forest. *J. Comput. Appl.* **2018**, *38*, 58–62.
- Asif, A.; Mukhtar, H.; Alqadheeb, F.; Ahmad, H.F.; Alhumam, A. An approach for pronunciation classification of classical Arabic phonemes using deep learning. *Appl. Sci.* **2022**, *12*, 238. [\[CrossRef\]](#)
- Cao, P. Identification and classification of Chinese traditional musical instruments based on deep learning algorithm. In Proceedings of the the 2nd International Conference on Computing and Data Science, Palo Alto, CA, USA, 28 January 2021; pp. 1–5.
- Zhang, K. Music style classification algorithm based on music feature extraction and deep neural network. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 1–7. [\[CrossRef\]](#)
- Alvarez, A.A.; Gómez, F. Motivic Pattern Classification of Music Audio Signals Combining Residual and LSTM Networks. *Int. J. Interact. Multi.* **2021**, *6*, 208–214. [\[CrossRef\]](#)
- Iloga, S.; Romain, O.; Tchuente, M. A sequential pattern mining approach to design taxonomies for hierarchical music genre recognition. *Pattern Anal. Appl.* **2018**, *21*, 363–380. [\[CrossRef\]](#)
- Wang, J.J.; Huang, R. Music Emotion Recognition Based on the Broad and Deep Learning Network. *J. East China Univ. Sci. Technol. Nat. Sci.* **2021**, 1–8. [\[CrossRef\]](#)
- Jia, N.; Zheng, C.J. Music theme recommendation Model based on attentional LSTM. *Comput. Sci.* **2019**, *46*, 230–235.
- Xia, Y.T.; Jiang, Y.W.; Li, T.R.; Ye, T. Deep Learning Network for The Classification of Beethoven's piano sonata creation period. *Fudan J. Nat. Sci.* **2021**, *60*, 353–359.
- Zhang, Z.; Chen, X.; Wang, Y.; Yang, J. Accent Recognition with Hybrid Phonetic Features. *Sensors* **2021**, *21*, 186258. [\[CrossRef\]](#) [\[PubMed\]](#)
- Mishachandar, B.; Vairamuthu, S. Diverse ocean noise classification using deep learning. *Appl. Acoust.* **2021**, *181*, 108141. [\[CrossRef\]](#)
- Lhoest, L.; Lamrini, M.; Vandendriessche, J.; Wouters, N.; da Silva, B.; Chkouri, M.Y.; Touhafi, A. MosAIC: A Classical Machine Learning Multi-Classifer Based Approach against Deep Learning Classifiers for Embedded Sound Classification. *Appl. Sci.* **2021**, *11*, 8394. [\[CrossRef\]](#)
- Farooq, M.; Hussain, F.; Baloch, N.K.; Raja, F.R.; Yu, H.; Zikria, Y.B. Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network. *Sensors* **2020**, *20*, 6008. [\[CrossRef\]](#)
- Li, W.; Li, Z.J.; Gao, Y.W. Understanding digital music—A review of music information retrieval technology. *J. Fudan Univ. (Nat. Sci.)* **2018**, *57*, 271–313.

28. Huang, X. Research on Opera Classification Method Based on Deep Learning. Master's Thesis, South China University of Technology, Guangzhou, China, 2020.
29. Huang, J.; Lu, H.; Meyer, P.L. Acoustic scene classification using deep learning-based ensemble averaging. In Proceedings of the the 4th Workshop on Detection and Classification of Acoustic Scenes and Events, New York, NY, USA, 25–26 October 2019; pp. 94–98.
30. Ba Wazir, A.S.; Karim, H.A.; Abdullah, M.H.L.; AlDahoul, N.; Mansor, S.; Fauzi, M.F.A.; See, J.; Naim, A.S. Design and implementation of fast spoken foul language recognition with different end-to-end deep neural network architectures. *Sensors* **2021**, *21*, 710. [[CrossRef](#)]
31. Liu, H. Analysis of Cantonese opera singing music. *Chin. Theatre* **2016**, *8*, 70–71.
32. Liu, Y.S.; Wang, Z.H.; Hou, Y.R.; Yan, H.B. A feature extraction method for malicious code based on probabilistic topic model. *J. Comput. Res. Dev.* **2019**, *56*, 2339–2348.
33. Zhu, H.; Ding, M.; Li, Y. Gibbs phenomenon for fractional Fourier series. *IET Signal Process.* **2011**, *5*, 728–738. [[CrossRef](#)]
34. Hasija, T.; Kadyan, V.; Guleria, K.; Alharbi, A.; Alyami, H.; Goyal, N. Prosodic Feature-Based Discriminatively Trained Low Resource Speech Recognition System. *Sustainability* **2022**, *14*, 614. [[CrossRef](#)]
35. Huang, G.X.; Tian, Y.; Kang, J.; Liu, J.; Xia, S.H. Long short term memory recurrent neural network acoustic models using i-vector for low resource speech recognition. *Appl. Res. Comput.* **2017**, *34*, 392–396.
36. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, MA, USA, 8–10 June 2015; pp. 1–9.