10th International Conference on Intelligent Control and Information Processing
Marrakesh, Morocco, December 14-19, 2019

Music Genre Classification Based on Chroma Features and Deep Learning

Leisi Shi
Software College
Xi'an Jiaotong University
Xi'an, China
1052165618@qq.com

Chen Li*
Software College
Xi'an Jiaotong University
Xi'an, China
63663529@qq.com

Lihua Tian
Software College
Xi'an Jiaotong University
Xi'an, China
5316238@qq.com

***Abstract*—Music genre classification is an important branch of content-based music signal analysis. It is a challenging task in the field of music information retrieval (MIR). At present, the method based on deep learning has achieved good results. This paper constructs a neural network framework for music genre classification based on chroma feature. The chroma feature can represent the time domain and the frequency domain of music character and consider the existence of harmony. Besides, it is independent of the timbre, volume, absolute pitch, which are completely irrelevant to the genre classification. It is relatively robust to the background noise and can represent the primary information such as monophonic and polyphonic music distribution. In this paper, we estimate the type of music audio based on chroma feature combined with deep learning network. We input this feature into VGG16 network for training, and improve the last three layers. In the experiment, the classifier is trained by GTZAN dataset. The experimental results show that the framework can obtain higher classification accuracy and better performance.**

***Index Terms*—music genre classification, deep learning, music information retrieval, convolutional neural network**

I. INTRODUCTION

Music genre classification is a very challenging but promising task in MIR (Music Information Retrieval) field. Music is an evolving art and there is no clear boundary between music styles, so automatic classification of music genres is a challenging issue. For massive music, traditional manual retrieval and classification methods are difficult to satisfy people. Music classification is essentially a pattern recognition problem, which mainly includes two aspects——feature extraction and classification. For this problem, many scholars have proposed different solutions.

Traditional music classification methods include rule-based classification methods, pattern matching methods, neighborhood searching methods, and hidden Markov model methods. In 2002, George Tzanetakis systematically proposed three features, which are representing timbre, rhythm and pitch, using Gaussian classifier, Gaussian mixture model (GMM) and K-near neighbor (KNN) classifier for experiments, and established the most commonly used and quite well known dataset including 10 genres——GTZAN [1]. In 2015, Zhuang Yan and others believed that the beat of music is an important semantic feature that reflected the style of different music genres. Therefore, he used the wavelet transformation to extract the low-frequency beat features of the music signal, combined with the acoustic features of MFCC, and his classification method achieved the accuracy with 68.73% overall on GTZAN dataset[2].

In recent years, more and more scholars have begun to apply deep learning to the field of music information retrieval. Honglak Lee et al. proposed a convolutional deep belief network (CDBN) to improve the accuracy of music genre recognition and singer recognition [4]. In 2015, Zhang Pengjing et al. proposed a convolutional neural network combined with K-maximization pooling to extract the invariant features of music, which achieved an average accuracy of 83.9% on GTZAN dataset [3]. Tom LH. Li et al. used convolutional neural network to extract musical pattern features. Experiments showed that convolutional neural networks can extract specific pattern features in different music with a small amount of prior knowledge [5]. Sander Dieleman et al. used the end-to-end learning method to extract and classify the original audio signal using convolutional neural network, and compared it with the method convolutional neural network based on spectral map. The result showed that the end-to-end learning can obtained AUC (area under curve) close to the result based on the convolution on the spectrogram [6]. In 2018, Hareesh Bahuleyan combined the hand-extracted features with the convolutional audio spectrum to obtain better accuracy [7].

In general, music genre classification has a great relationship with harmony, but it has no relationship with human timbre, volume, absolute pitch, etc. These irrelevant factors greatly affect the improvement of classification accuracy, thus leading to the accuracy of existing algorithms are all not above 90%. Therefore, how to ignore these factors in our algorithm and focus on the important factors that have a crucial impact on the classification of music genres is our main task. Based on this consideration, we construct a framework for music classification using chroma features in this paper. This classification framework can consider the existence of harmony and ignore the information which is irrelevant with genre classification such as absolute pitch, volume and velocity. It can also represent the primary information such as the scale distribution of monophonic and polyphonic music, so it is relatively robust to background noise, and has achieved better results.

The structure of this paper is as follows. In section 1, the key points that can be improved in previous research are analyzed, and the feasibility of chroma features is summarized. In section 2, we introduce the extraction method of features. In section 3, we introduce our feature extraction

This work is supported by the National Natural Science Foundation of China under Grant No. 61901356 and the HPC Platform of Xi'an Jiaotong University.

method and network architecture. In section 4, the experimental process, final results and the comparative analysis are showed. In section 5, we draw the final conclusion and introduce the future work.

II. Chroma Feature

The chroma feature is a mainstream audio melody feature in the field of music information retrieval. It is designed according to Twelve Tone Equal Temperament. Since, in music, notes exactly one octave apart are perceived as particularly similar, knowing the distribution of chroma even without the absolute frequency (i.e. the original octave) can give useful musical information about the audio, and may even reveal perceived musical similarity that is not apparent in the original spectra. Chroma feature is usually represented as a 12-dimensional vector v=[V(1), V(2), V(3),...V(12)], and each element of a vector is associated with one element of the set of {C,C#, D, D#, E, F, F#, G, G#, A, A#, B}, reflecting the local energy distribution of the audio signal at semitones represented by the 12 pitch names [8].

Each element of Chroma feature vector can be calculated by the equation (1)

$$Chroma(n) = \sum_{i=1}^{nPeaks} w(n, f_i) \cdot a_i^2, n = 1,2, \cdots 12\,, \quad (1)$$

where, a_i and f_i are the amplitude and frequency of the i-th peak of the signal, respectively. $w(n, f_i)$ is the weight of the signal of frequency f_i for the semitone n. It is calculated as following steps.

First, calculate the central frequency f_n of each semitone by the equation (2)

$$f_n = f_{rep} \cdot 2^{\frac{n}{12}}, n = 1,2, \cdots 12, \quad (2)$$

where f_{rep} is the fundamental frequency of original music signals.

The musical interval distance between each peak of signals and the central fundamental frequency f_n of the semitone is defined as the equation (3)

$$d = 12 \cdot \log_2 \left[\frac{f_i}{f_n}\right] + 12 \cdot m\,, \quad (3)$$

where m is an adjustment factor for an integer value in order to ensure the value of $|d|$ is the smallest. Then we can calculate the weight according to the equation (4)

$$w(n, f_i) = \begin{cases} -\cos^2\left(\frac{\pi}{2} \cdot \frac{d}{0.5 \cdot l}\right), & |d| \leq 0.5 \cdot l, \\ 0, & |d| > 0.5 \cdot l, \end{cases} \quad (4)$$

where l is the pre-set length of the weighted window [11].

III. The Construction of Music Classification Framework Based on Chroma Features

A. Extraction of CQT-Based Chroma Feature

The chroma feature can be obtained by CQT (hereafter referred to as chroma_cqt) [11], or by STFT (hereafter referred to as chroma_stft) [9][10]. Both features are available in the audio field. "Twelve Tone Equal Temperament" is common internationally used. In this architecture, 12 pitches of each octave are arranged in ascending ordered by the interval of a semitone, and the frequency ratio of two notes with one octave apart is 2:1. It is obvious that the frequency ratio of two notes with a semitone apart is 2^(1/12):1. Therefore, the scale frequencies of music signals are distributed exponentially. However, when FFT or STFT is used to estimate the frequency of musical scale, the spectral line of frequency is linearly distributed, so the frequencies of the two above cannot correspond to each other completely, causing errors in estimating the frequencies of certain scales. Therefore, in the phase of time-frequency transformation, this paper will adopt a time-frequency transformation method, CQT, which has the same exponential distribution of spectral line frequency and scale frequency, to achieve better performance.

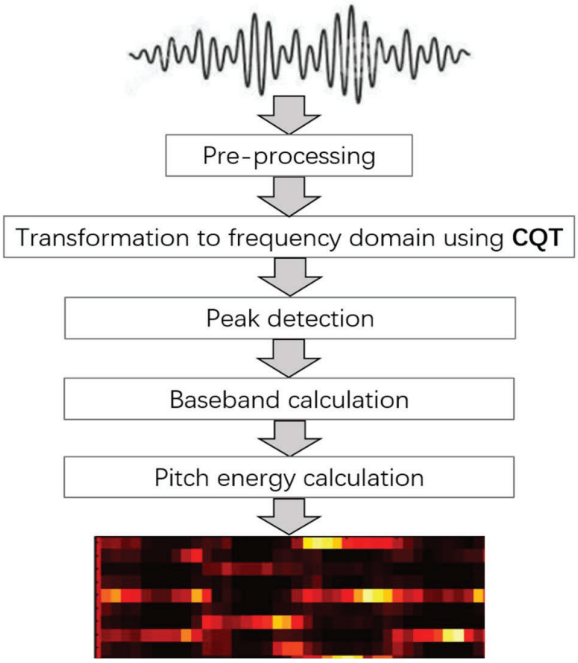

Fig. 1. Extraction process of chroma features.

In general, the extraction process of chroma feature is shown as Fig. 1.

In the preprocessing stage of feature extraction, it is necessary to convert the original audio input into a single channel signal with a fixed sampling rate. Then, the processed signal is subdivided into several frames. In this stage, a continuous and partially overlapping (eg, 50%) short-term frame (eg, 100 ms) is usually adopted. And then, these frames is computed one by one through a constant Q transform (CQT). Finally, the spectrum of the signal can be obtained.

The calculation process of CQT is shown in equation(5)

$$f_k = 2^{\frac{k}{\beta}} f_{min}, \quad (5)$$

where f_k refers to the frequency value of the k-th component, and f_{min} is the lower limit frequency of the processed signal. β is the number of frequency lines contained in an octave. For example, β=36 represents 36 frequency lines per octave. That is, there are three frequency components in each semitone range. And f_k and f_{k+1} satisfy the following equation (6)

$$f_{k+1} - f_k = f_k \frac{1}{\frac{1}{a^{\beta}} - 1}. \quad (6)$$

Let δ_f be the bandwidth at frequency f, also known as frequency resolution. According to the definition, CQT is a transformation whose frequency to bandwidth ratio is constant Q, and Q satisfies equation (7)

$$Q = \frac{f}{\delta_f} = \frac{1}{a^{\frac{1}{\beta}}-1}. \quad (7)$$

Obviously Q is a constant determined by β.

Note that N_k is the window length varying with frequency, and f_s is the sampling rate. We get the equation (8)

$$N_k = \frac{f_s}{\delta_{f_k}} = \frac{f_s \cdot Q}{f_k}. \quad (8)$$

According to the constant Q, CQT obtains different frequency resolutions by adopting different window lengths, thus obtaining the frequency value of each semitone. The frequency component of the k-th semitone of the n-th frame in CQT can be expressed as the equation (9)

$$X_n^{cqt}(k) = \frac{1}{N_k}\sum_{m=0}^{N_k-1} x(m) w_{N_k}(m) e^{\frac{-j2\pi mQ}{N_k}}, \quad (9)$$

where $x(m)$ is a time domain signal and $w_{N_k}(n)$ is a window function with the window length N_k [11].

Through CQT transformation, we get the amplitude spectrum of the signal, and then find out the maximum n peaks in the fixed frequency range in the spectrum, and calculate the difference among them. According to the A4 standard frequency 440Hz specified in the Twelve Tone Equal Temperament, the fundamental frequency f_{rep} of the original music signal can be inferred. Finally, the chroma_cqt features are obtained according to the flow chart Fig. 1 and formula (1).

B. Construction of Neural Networks

The audio signal is one-dimensional time domain signal with a complex waveform, so it is difficult to observe the characteristics of the audio intuitively. Although the commonly used RNN model has better performance in time domain of audio information, its processing ability in the frequency domain is weak. If only the frequency domain information of audio signal is processed, the time domain information of audio would be lost. Based on the above considerations, the acoustic spectrogram contains two dimensions of information: time domain and frequency domain. Therefore, combining with the above analysis, this paper considers the use of chroma-based spectrogram for music analysis and classification. Convolutional neural network has a very outstanding ability in image classification, so we will use the spectrogram based on the chroma feature as the input of convolutional neural network, train the network and finally get the music classification results.

VGG16 network has a strong feature learning ability. It has a large number of trained parameters and weights. In particular, the convolution layer has a strong ability to extract features such as image edges, contours, content and others. So this paper will adopt the network structure of VGG proposed by Simonyan and Zisserman in 2014 [15], and make appropriate adjustments according to the characteristics of music classification.

Because there are 4096 parameter nodes in the dense layer of the original network, accounting for 70% of the total parameter nodes, which seriously affects the processing efficiency of the network, at the same time, since the amount of audio information is less than that of pictures, 4096 nodes are prone to overfitting during the training process. Therefore, we add another dense-layer with the units of 256 and a dropout-layer with the rate of 0.5 to avoid overfitting.

After the processing with convolution and pooling for the chroma feature, the result is integrated into a one-dimensional array through the softmax function. A probability distribution P based on the data input is obtained, and finally 10 music genre labels are obtained according to P.

The network structure is shown in Fig. 2.

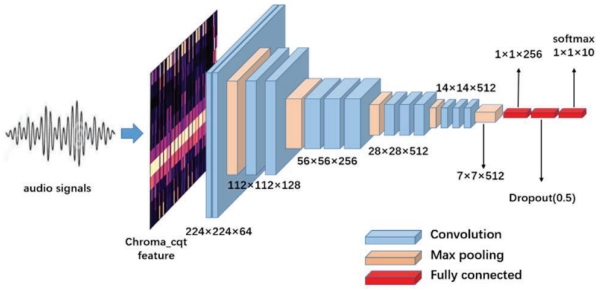

Fig. 2. Neural network structure.

The VGG16 convolutional neural network is consisted of 16 weighting layers, which are 13 convolutional layers and 3 dense layers. The input of network is an image of 224 pixels × 224 pixels × 3 vectors. Usually 3 × 3 filters in the convolution layer are used, and then continuous stacking of 2 or 3 filters constitutes a convolution sequence for the effect of a larger receptive field. The sliding step size is 1. The boundary filling technique is used to keep the data dimension unchanged. The pooling layer uses a 2×2 pooling window with a step size of 2, which is used to reduce the size of the feature image after convolution and to ensure the translation invariance of the model.

The network input is chroma feature with a feature dimension of 19000×m×129×3, where m represents the value of the frequency resolution, and the data is through the convolution layer to output multiple feature maps. The convolution kernel is a weighted matrix for convolution. When the convolution kernel moves according to the set step size, the data of the corresponding position in the input is weighted and summed to obtain the output value of the corresponding position in the feature map.

The calculation equation for convolution is as the equation (10)

$$f_{i,j} = h\left(\sum_{m=0}^{F_w-1}\sum_{n=0}^{F_H-1} w_{m,n} x_{i+m,j+n} + b\right), \quad (10)$$

where $f_{i,j}$ is the value of the i-th row and j-column in the feature map, w is the weighted matrix of the convolution kernel, x is the input matrix, b is the offset of the convolution, and h is the activation function of the convolution layer. F_w and F_H are the width and height of the sum of convolution. It can be seen that each output value is computed by the convolution kernel and the local part of the input data, and one feature map is calculated by one convolution kernel. This is the embodiment of local connection and weight sharing, which greatly reduces the number of parameters.

After convolution, the relationship between the resulting feature map and the input is as the equation (11) and (12)

$$W_F = \frac{W - F_W + 2P_W}{S_W} + 1, \quad (11)$$

$$H_F = \frac{H - F_H + 2P_H}{S_H} + 1, \quad (12)$$

where W_F and H_F are the width and height of the feature map, W and H are the width and height of the input, P_W and P_H are the filling sizes of the input data in the width direction and the height direction, S_W and S_H are step size of the sum of convolution in the width and height directions.

TABLE I. DIMENSION IN DIFFERENT FREQUENCY RESOLUTIONS

Frequency resolution	***36***	***48***	***60***	***72***
Feature dimension	19000×36×129×3	19000×48×129×3	19000×60×129×3	19000×72×129×3

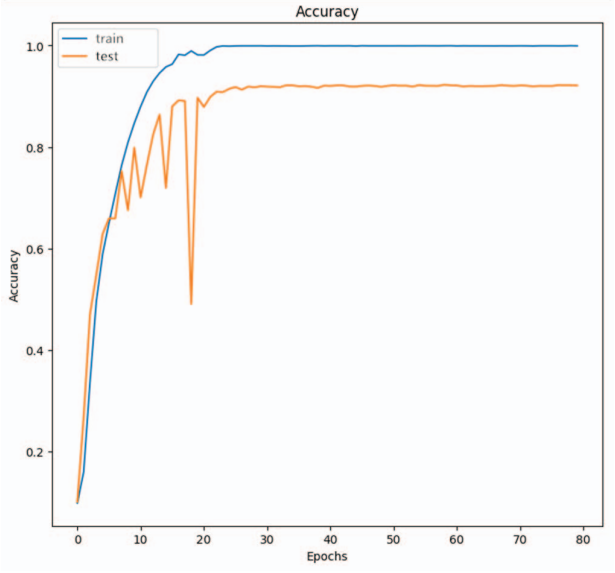

Fig. 3. Accuracy curve with frequency resolution 48.

The pooling layer is used to downsample the feature maps, compress the size of the feature map, reduce the amount of calculation, and increase the generalization ability of the model, and enhance the robustness. We choose the max-pooling in this experiment. The output gradient δ_l of the network at l-th layer is returned to the position corresponding to the maximum value of the l-1 layer, and the other areas are 0.

The dense layer is used to get the final 10 classification labels as the equation (13)

$$P(j) = \mathrm{P}\left[L = l_j | a; (w, b)\right], \tag{13}$$

where l_j represents j categories of the tags, and the optimal model for the convolutional neural network is the model whose loss function $L = (w, b)$ is the smallest (this paper selects the cross entropy function).

IV. EXPERIMENTS AND RESULTS

A. Dataset

Our experiment is based on the dataset named GTZAN, which was collected by G. Tzanetakis and P. Cook in 2000-2001. It is a widely used and recognized dataset for music classification.

The audios in GTZAN are collected from a variety of sources including radio, personal CDs, microphone recordings, in order to represent a variety of recording conditions.

The dataset is consisted of 1000 songs with 30 seconds long. It contains 10 genres(blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, rock) and each genre is represented by 100 songs. The songs are all 22050Hz mono 16-bit audio files [1].

TABLE II. ACCURACY UNDER DIFFERENT FREQUENCY RESOLUTIONS

Frequency resolution	**Chroma_cqt**			
	36	***48***	***60***	***72***
Accuracy	91.64%	92.12%	91.67%	90.92%

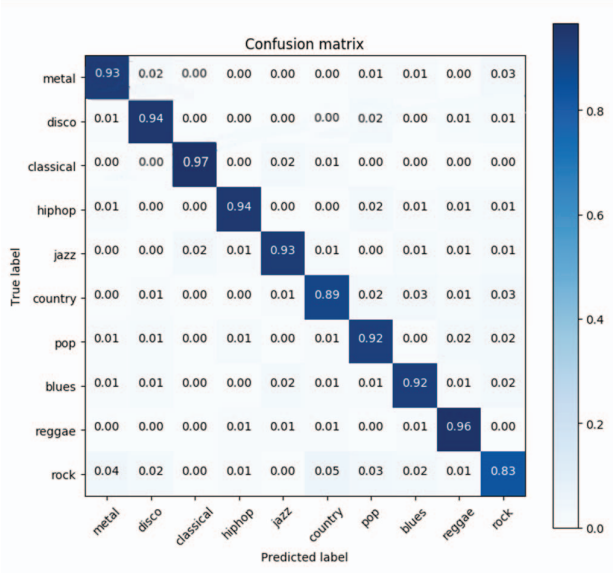

Fig. 4. Accuracy curve with frequency resolution 48.

B. Experimental Parameters

We first segment the audio into several frames by 3s long and 50% overlap, and then extract the chroma features of the signal through the CQT transformation at the sampling rate of 22050. Because the performance of the chroma feature is related to the frequency resolution selected, we extract the chroma features of 36, 48, 60, and 72-dimensional frequency resolutions respectively(the default frequency resolution of the chroma feature is 12). At different frequency resolutions, the input dimensions of the data are shown in the TABLE I.

Then we divide the dataset into 70% training set and 30% testing set. Our network directly uses the already trained parameters of ImageNet as the established parameters. The others are as follows:

Optimizer=SGD
Batch_size=8
Epoch=100

Finally, according to the existing classification criteria, the results are divided into 10 categories, including blues, classic, country, disco, hiphop, jazz, metal, pop, reggae, rock.

C. Experimental Results and Analysis

Based on the above datasets and related parameters, the experimental results of this framework at different frequency resolutions are shown as TABLE II.

It can be seen from TABLE II that the best accuracy is obtained when the frequency resolution is 48. The accuracy curve and confusion matrix at this time are shown in Fig. 3 and Fig. 4 respectively.

It can be seen that there is no completely positive correlation between frequency resolution and classification results. The frequency resolution refers to the minimum interval of distinguishing two different frequency signals. The frequency resolution should coincide with the minimum interval of different frequencies of the original audio signal. Only in this way, can we get the best sampling results and

TABLE III. THE COMPARISON RESULT WHEN USING DIFFERENT CHROMA FEATURES

Frequency resolution	***36***	***48***	***60***	***72***
Accuracy of chroma_cqt	91.64%	92.12%	91.67%	90.92%
Accuracy of chroma_stft	85.05%	86.30%	86.32%	86.28%

feature extraction results. If the frequency resolution is too large, it would result in multiple sampling in one frequency cycle, then the sampling features would not be representative. And if the frequency resolution is too small, the sampling results would be too rough.

In addition, in order to prove that chroma_cqt has better effect than chroma_stft, we extract the features of chroma_cqt and chroma_stft separately, and then we input two features into the neural network for comparison and observe the final result in this paper. The other experimental parameters are the same as above.

The comparison results of the classification accuracy when using different chroma features are shown as TABLE III.

From TABLE III, chroma_cqt achieves a better performance than chroma_stft. The possible reason lies in that the musical signal scale frequency is distributed exponentially, but when STFT is used to estimate the frequency of musical scale, the spectral line of frequency is linearly distributed, so the frequencies of the two above cannot correspond to each other completely, causing errors in estimating the frequencies of certain scales. However, the transformation method CQT has an exponential distribution as same as the scale frequency, so chroma_cqt can achieve better performance.

To further demonstrate the performance of our method, TABLE IV shows the comparison result of our method and other current methods， whose datasets are all based on GTZAN.

From TABLE IV, it can be seen that by inputting chroma_cqt features into the improved VGG16 network, the classification accuracy of the GTZAN dataset can reach the highest 92.12%, far exceeding the other algorithms. In the stage of feature extraction, we fully consider the important influence effects of harmony on the music genre, and exclude the influence factors such as absolute pitch, volume and intensity, which are irrelevant to the genre classification. At the same time, the music spectrum combined with the frequency domain information and time domain information can ensure the completeness of feature information. Finally, based on the improved version of VGG16, which has a very good classification effect, the music genre is classified. Other methods all do not exclude the above irrelevant factors, which would hinder the improvement of classification accuracy.

V. CONCLUSIONS AND FUTURE WORK

Music genre classification has a great relationship with harmony, but has no relationship with human timbre, volume, absolute pitch, etc. These irrelevant factors greatly affect the improvement of classification accuracy. Therefore, how to ignore these factors in the algorithm and focus on the kernel factors that have a crucial impact on the classification of

TABLE IV. THE COMPARISON RESULT OF CLASSIFICATION ACCURACY BETWEEN OUR METHOD AND OTHERS

Method	***Our method***	***Reference [14]***	***Reference [13]***	***Reference [12]***
Year	2019	2018	2018	2017
Accuracy	92.12%	89.00%	87.00%	75.50%

music genres is our main task. Based on this consideration, we construct a framework for music classification using the chroma feature and improved VGG16 deep learning network in this paper. This classification framework considers the existence of harmony and ignores the information which is irrelevant with genre classification such as absolute pitch, volume and velocity. It can also represent the primary information such as the scale distribution of monophonic and polyphonic music, so it is robust to the background noise. Moreover, the classification accuracy of the chroma_cqt is much higher than that of chroma_stft. The highest classification accuracy of this model can reach 92.12%, far exceeding that of all previous models.

During the experiment, we have tried to use Resnet50 neural network for training, but the experimental accuracy was less than 75%. Therefore, the follow-up work of this experiment will focus on analyzing the influence of network structure on music genre classification and verifying it on different datasets in order to find the most suitable network structure for music genre classification.

REFERENCES

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," IEEE Trans. Speech Audio Process, vol. 10, no. 5, pp. 293-302, July 2002.

[2] Y. Zhuang and F. Yu, "Combining beat semantic features and MFCC acoustic features for music genre classification," CE&A, vol. 51, no. 3, pp. 197-201, 2015.

[3] P. Zhang, X. Zheng, W. Zhang, and S. Li, "A Deep neural network for modeling music," in Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ACM, 2015, pp. 379-386.

[4] H. Lee, P. Pham, Y. Largman, and A. Ng, “Unsupervised feature learning for audio classfication using convolutional deep belief networks,” Adv Neural Inf Process Syst, pp. 1096–1104, 2009.

[5] T. Li, A. Chan, and A. Chun, “Automatic musical pattern feature extraction using convolutional neural network,” in Proc. Int. Conf. Data Mining and Applications, 2010.

[6] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, 2014, pp. 6964-6968.

[7] A. Karatana and O. Yildiz, "Music genre classification with machine learning techniques," in 2017 25th Signal Processing and Communications Applications Conference (SIU), Antalya, 2017, pp. 1-4.

[8] X. Zhang, N. Li, and W. Li, "Verification for robustness of chroma feature," Computer Science, 2014.

[9] M. Bartsch and G. Wakefield, "Audio thumbnailing of popular music using chroma-based representations," IEEE Trans. Multimedia, vol. 7, no. 1, pp. 96-104, Feb. 2005.

[10] D. Ellis and G. Poliner, "Identifying `cover songs' with chroma features and dynamic programming beat tracking," in 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, Honolulu, HI, 2007, pp. IV-1429-IV-1432.

[11] J.Brown, "Calculation of a constant Q spectral transform," J. Acoust. Soc. Am., vol. 89, no.1, pp. 425, 1991.

[12] R. Rajan and H. Murthy, "Music genre classification by fusion of modified group delay and melodic features," in 2017 Twenty-third National Conference on Communications (NCC), Chennai, 2017, pp. 1-6.

[13] S. Sharma, P. Fulzele, and I. Sreedevi, "Novel hybrid model for music genre classification based on support vector machine," in 2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE), Penang, 2018, pp. 395-400.

[14] P. Fulzele, R. Singh, N. Kaushik, and K. Pandey, "A hybrid model for music genre classification using LSTM and SVM," in 2018 Eleventh International Conference on Contemporary Computing (IC3), Noida, 2018, pp. 1-3.

[15] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in International Conference on Learning Representation, San Diego, 2015, pp. 1-15.