J Inf Process Syst, Vol.20, No.1, pp.13~23, February 2024
https://doi.org/10.3745/JIPS.04.0300

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Musical Genre Classification Based on Deep Residual Auto-Encoder and Support Vector Machine

Xue Han[1], Wenzhuo Chen[2], and Changjian Zhou[2,*]

Abstract

Music brings pleasure and relaxation to people. Therefore, it is necessary to classify musical genres based on scenes. Identifying favorite musical genres from massive music data is a time-consuming and laborious task. Recent studies have suggested that machine learning algorithms are effective in distinguishing between various musical genres. However, meeting the actual requirements in terms of accuracy or timeliness is challenging. In this study, a hybrid machine learning model that combines a deep residual auto-encoder (DRAE) and support vector machine (SVM) for musical genre recognition was proposed. Eight manually extracted features from the Mel-frequency cepstral coefficients (MFCC) were employed in the preprocessing stage as the hybrid music data source. During the training stage, DRAE was employed to extract feature maps, which were then used as input for the SVM classifier. The experimental results indicated that this method achieved a 91.54% F1-score and 91.58% top-1 accuracy, outperforming existing approaches. This novel approach leverages deep architecture and conventional machine learning algorithms and provides a new horizon for musical genre classification tasks.

Keywords

Deep Residual Auto-Encoder, MFCC, Music Artificial Intelligence, Musical Genre Classification

1. Introduction

Massive music data is produced worldwide every day and plays an irreplaceable role in entertainment. People can choose different musical genres based on their mood. In certain situations, a particular genre of music is needed, such as in cafés, where jazz is more popular while pop music is more suitable for playgrounds. However, identifying musical genres using massive amounts of music data is time-consuming and laborious. In recent years, there have been two main approaches to classifying musical genres: one involves manually extracting music features for classification, while the other involves using automatic machine learning tools for musical genre classification. The frequency domain feature is a common method in audio data processing and has been adapted for music genre classification for a long time [1]. Spectral features such as pitch, timbre, audio frequency, and tonality are depicted in the frequency domain and are important for music genre classification. Chord transition features were introduced to identify Western music tonals, offering a novel method [2]. Foleis and Tavares [3] proposed a method for extracting sound texture features for music genre classification, which involves using linear

Manuscript received December 20, 2021; first revision March 29, 2022; second revision April 29, 2022; accepted May 25, 2022.

* **Corresponding Author:** Changjian Zhou (zhouchangjian@neau.edu.cn)
[1] School of Arts, Northeast Agricultural University, Harbin, China (hanxue@neau.edu.cn)
[2] Dept. of Data and Computing, Northeast Agricultural University, Harbin, China (chenwenzhuo@neau.edu.cn; zhouchangjian@neau.edu.cn)

downsampling as a texture feature. The music texture selector, based on K-means, was designed to extract the sound texture features of each track and achieved satisfactory results. With the advancement of computer vision technology, researchers have explored visual features and evaluated their feasibility for music genre classification tasks. Vidwans et al. [4] exploited the melodic contours and dynamic variations of music and extracted melodic features, which served as the most intuitive feature representations for classification. With the widespread implementation of machine learning and deep-learning architectures in various fields, numerous studies have attempted to classify musical genres using machine-learning models. Puppala et al. [5] proposed a method for music-genre identification based on a convolutional neural network (CNN). The novel method adopted a CNN as an encoder for feature extraction and input into machine learning classifiers for identification, which achieved high accuracy. Chen and Steven [6] proposed a method that combines transfer and active learning to classify musical genres. The novel method achieved higher accuracy than conventional machine learning approaches. Conceicao et al. [7] analyzed the characteristics of Brazilian music and concluded that it mainly consists of pop and rock styles. A supervised learning method was used for classification, using major databases such as AudioSet and GTZAN. This demonstrates the efficiency of the Brazilian musical classification process. Pelchat and Gelowitz [8] analyzed recent developments in musical genre classification methods and concluded that limited success has been achieved because of the various formats, databases, and music types. They proposed an efficient method for classifying musical genres utilizing neural networks. The deep learning method is employed for audio signal processing, and the combination of deep CNN and handcrafted features is a novel method for musical genre classification [9]. As a variety of deep CNNs have been presented in the field of signal processing, it has become possible to create a robust and efficient model for musical genre classification.

Existing studies provide valuable insights into the analysis of musical genres. However, they have the following limitations: it is impracticable to manually extract music features from large amounts of music data during the feature extraction stage. Furthermore, meeting the actual requirements of accuracy and timeliness is challenging. In this study, we proposed a combined method of a deep residual auto-encoder (DRAE) and support vector machine (SVM) and evaluated its feasibility for classifying musical genres. The contributions of this study are as follows:

- A novel hybrid machine learning method that combines the DRAE and SVM was proposed for musical genre classification and achieved excellent performance.
- Two main streams of features were obtained, comprising eight traditional music features and 20 Mel-frequency cepstral coefficient (MFCC) features, which were combined to form the initial input vectors for classification.

The remainder of this paper is organized as follows: the related works are presented in Section 2. The proposed method is detailed in Section 3. Section 4 describes the experiment and analysis, while Section 5 presents the discussion. Section 6 presents the conclusion.

2. Related Works

This work presented a novel deep learning encoder method and CNN combined approach for musical genre classification approach. The novel method is closely related to three branches of this study, such as MFCC, DRAE, and SVM. A brief review that leads to the proposed methodology is given as follows.

2.1 Mel-Frequency Cepstral Coefficients

Feature engineering is an irreplaceable role in machine learning process, appropriate features can condense samples and highlight the most representative feature representation for machine learning models. As one of the most used audio feature extraction tools, MFCC is introduced for combining human auditory perception and the speech generation mechanism, which can describe the generated phoneme accurately [10]. The MFCC architecture is illustrated in Fig. 1 which mainly curated into the following stages.

1) Pre-emphasis, framing and windowing for inputted audio signal. In fact, pre-emphasis is a high-pass filter as (1):

$$s_p(n) = s(n) - \lambda s(n-1), \ \ \lambda \in [0,1] \tag{1}$$

where n is the length of signal frame, $s(\cdot)$ is the original signal, and λ is a constant. The purpose of pre-emphasis is emphasizing high-frequency part, removing lip radiation influence, increasing speech high-frequency resolution, and making signal spectrum flat to keep it in the whole frequency band from low frequency to high frequency, as well as calculate the spectrum with the same signal-to-noise ratio. As the change of speech signal is rapid while Fourier transform is usually suitable for analyzing stable signals, it is necessary to frame the long signal into short frames to make it convenient for calculation. To reduce the sidelobe intensity, the Hamming window is introduced by multiplying between each frame signal and smooth window function. A window function (the width of the window function is the frame length) should be selected for each frame.

2) Fast Fourier transform (FFT): It is hardly to classify signal features from the transformation in time domain, and therefore the signals are usually transformed into the energy distribution in the frequency domain while different energy distributions represent the different characteristics of signal. To address this issue, a spectrum with energy distribution is obtained by FFT when multiplied by window function.

3) Mel spectrum analysis: Mel spectrum analysis is based on human auditory perception, which focuses on specific frequency components by allowing certain frequency signals and ignoring the frequency signals that are apathetic to perceive as depicted in (2).

$$mel(f) = \mu \log{(1 + f/\sigma)} \tag{2}$$

where μ and σ are constants, f is the filter.

4) Mel cepstral analysis: To obtain MFCC features, a Mel filter $\log{(mel(f))}$, the discrete cosine transform (DCT) was adopted for calculating MFCC.

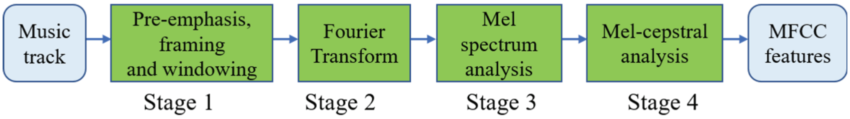

Fig. 1. The architecture the MFCC.

2.2 Convolutional Auto-Encoder

Convolutional auto-encoder (CAE) is inspired by auto-encoder by adding convolution layers for feature encoding [11] (Fig. 2). As one of the most representative machine learning models, auto-encoder adopts unsupervised learning and feature representation approaches, and it shines brightly in academia and industry. Two main components like encoder unit and decoder unit are embedded in the model (encoder process) while encoder unit maps input samples into feature space [12], and decoder unit maps feature to original space. The convolution layer can be demonstrated as (3).

$$C = \tau\left(\sum h^x \Delta \widehat{M}^x + c\right), \qquad h^x = \tau(\chi \varDelta M^x + b^x) \tag{3}$$

where Δ denotes convolution operation, $\widehat{M}^x$ is the rot 180° of M^x, τ is the constant coefficient, b is the convolution bias matrix.

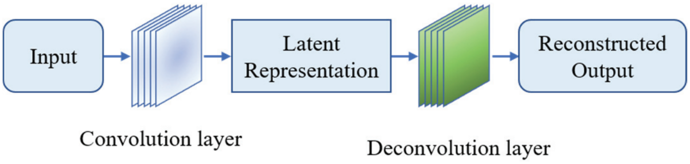

Fig. 2. The architecture of convolutional auto-encoder.

2.3 Support Vector Machine

As one of the widely used machine learning models, SVM achieved unparalleled success in various fields with its strong and powerful capability of feature representation [13]. The SVM algorithm finds a maximum hyperplane among different categories when given limited samples [14]. SVM is rarely disturbed by noise and states wonderful performance [15,16]. Assume y denotes the sample labels and $X = \{x_1, x_2, x_3, \dots, x_n\}$ denotes the samples, all the samples obey the following formula (4):

$$y \cdot (M^T \cdot x_i + \lambda) - 1 \geq 0 \tag{4}$$

where M is the hyperplane matrix and is the constant parameter.

SVM aims to find the maximum value of d to make the distance of different types of hyperplanes the farthest as depicted in Fig. 3. There has been a lot of literature on this topic, and the calculation methods will not be repeated here.

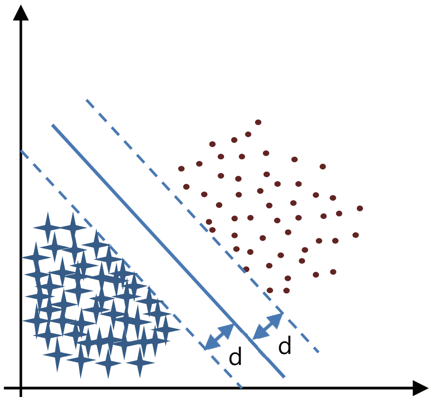

Fig. 3. The SVM architecture.

3. Proposed Method

CAE and SVM have achieved significant success in the field of computer vision. To evaluate their feasibility for musical genre classification, a hybrid machine learning model that combines a DRAE and SVM was proposed. The architecture of the proposed approach is shown in Fig. 4. The main components, such as feature fusion and the DRAE, are detailed as follows.

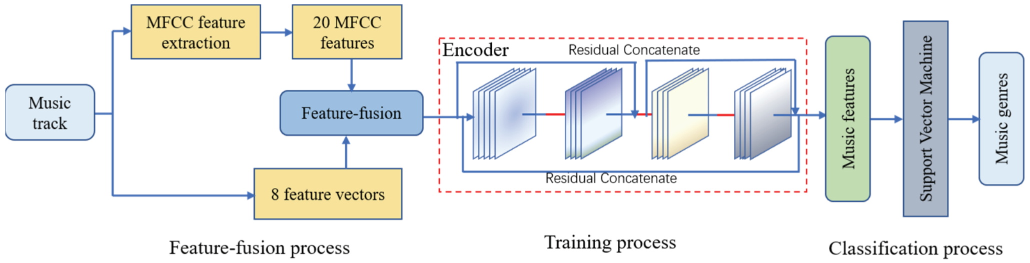

Fig. 4. Architecture of the proposed method.

3.1 Feature Fusion

In this study, we present a novel multi-feature fusion-based approach for extracting music features. Eight conventional musical features, including audio root mean square (RMS) power, spectral centroid, spectral bandwidth, spectral roll-off, poly features, zero-crossing rate, tempogram, and Tonnetz, are manually extracted from the input samples. A brief overview of these features is provided in Table 1.

In this study, MFCC was employed to extract musical features from an input audio sequence. The MFCC parameters, along with their first-order differentials, were returned. The Mel filter order was set to 20, the length of the FFT was 256, the sampling frequency was 10,000 Hz, and the 256 points were divided into frames. A total of 20 audio MFCC features were extracted. Fused with the previous eight audio features, they were combined as the preprocessed music features. In this study, 12,400 samples from ten musical genres were used for the experiment.

Table 1. Main adopted features of music

Feature	Introduction
Audio RMS power	It is used to record the amount of audio energy
Spectral centroid	It is used for measuring signal distribution of the frequency components
Spectral bandwidth	It is a measure of the certain number of decibels down from the spectral maximum
Spectral roll off	It is a measure of spectrum skewness
Poly features	Get coefficients of fitting an nth-order polynomial to spectrogram columns
Zero crossing rate	It is the sign-change rate along a frame and is used to characterize percussive sounds and environmental noise
Tempogram	Local onset strength envelope autocorrelation
Tonnetz	Computes the tonal centroid features

3.2 Deep Residual Auto-Encoder

A residual connection was introduced in the CAE, as shown in Fig. 5, since residual connection-based networks have been proven to prevent gradient degradation and achieve better convergence performance for image recognition [17]. This study aimed to investigate the adaptability of residual connections in audio feature encoding. The encoder unit consisted of one global and two local residual connections. Batch normalization and ReLU activation functions were introduced after each layer of convolutions. The concept of combining global and local residual connections aims to maintain gradients without degradation, thereby preserving the model's ability to converge effectively. When considering the input samples as $X = \{x_1, x_2, x_3, \dots, x_n\}$ and the output feature vectors as y, the residual block in the DRAE was defined as (5).

$$y = x + f(x_i, \{R_i\}) \tag{5}$$

where R_i denotes the residual mapping of x_i. In this study, the fused features were input into the DRAE for encoding, and the output feature mapping of the DRAE was input into the SVM for classification.

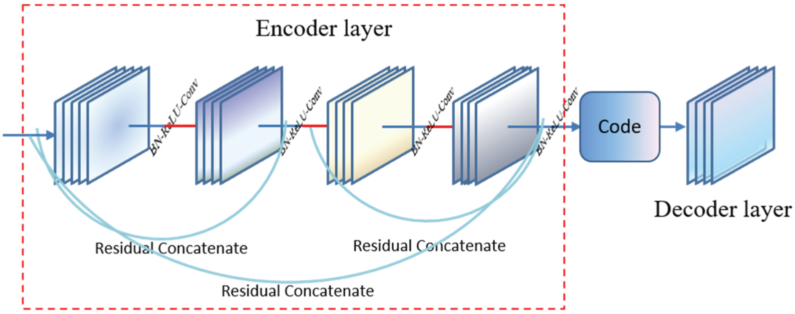

Fig. 5. Deep residual auto-encoder.

4. Experiment and Analysis

4.1 Data Acquisition

A total of 12,400 music tracks were collected in this study: 1,000 from the public CTZAN dataset [18], 9,400 from QQ music, and 2,000 from NetEase cloud music. All musical genres were labelled as classical, country, edm_dance, jazz, kids, Latin, metal, pop, R&B, and rock. The beat per minute (BPM) for each genre is demonstrated in Fig. 6. The dataset was curated into three classes: 70% for training, 15% for validation, and 15% for testing.

4.2 Training Details

In this study, the music data was first extracted from eight traditional music features and 20 MFCC features, and then the two categories were combined into fused features. All fused music data features were input into the DRAE for feature extraction, and SVM was employed for classification. The hyperparameters are listed in Table 2.

Loss function

In this study, the cross-entropy loss function was employed to measure the difference between the target and predicted values, which can be demonstrated as (6).

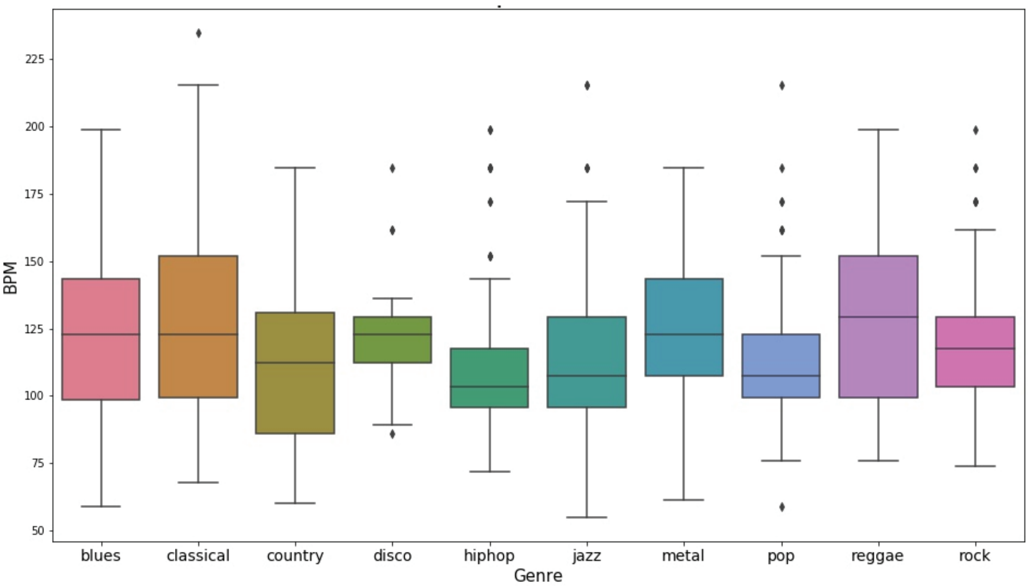

Fig. 6. BPM of music genres.

$$\text{Cross} - \text{entropy} = E_{x\sim p}[-\log P(x)] \tag{6}$$

where $P(x)$ is the probability distribution of the variable x and $E_{x\sim p}$ denotes the expectation of the x probability distribution. The Cross-entropy function was obtained under the model constraints, making the output distribution of the model close to the distribution of the training data.

Optimizer function

The root mean square propagation (RMSProp) optimizer function is an improvement of AdaGrad algorithm, which introduces the weight parameters   to the weight sum of the corresponding component. The RMSProp optimizer function is obtained as (7).

$$Sdx_i = \rho Sdx_i + (1-\rho)(dx_i)^2 \tag{7}$$

where x is the horizontal-axis direction parameter, ρ is the attenuation coefficient, and $(dx_i)^2$ is the square of dx_i.

Kernel function

The SVM classifier was introduced in this study for musical genre classification, and the DRAE extracts features as input into the SVM for music genre identification. The radial basis function (RBF) was utilized as the kernel function, as shown in (8).

$$kernel(x) = e^{(-\frac{\|x-c\|^2}{2k^2})}, \qquad k > 0 \tag{8}$$

where k is a constant parameter and is the coordinate of the center point of the hyperplane. In this study, the RBF is employed to address multivariable interpolation to adapt to multi-classification problems.

Table 2. Hyperparameters

Parameter	Value
Loss function	Cross-entropy
Optimizer function	RMSProp
Activation function	ReLU
Initial learning rate	0.0001
SVM kernel function	Radial basis function

4.3 Results and analysis

Evaluation metrics such as precision, F1-score, specificity, and accuracy were introduced in this experiment to evaluate the existing and proposed models. In this study, conventional machine learning algorithms such as ResNet-50 and DenseNet-121 were employed for training and testing. In addition, recently published methods for classifying musical genres, such as those developed by Foleis and Tavares [3], Puppala et al. [5], Singh and Biswas [19], Yu et al. [20], and Liu et al. [21], were adopted for comparison using the same dataset. Conventional machine learning models typically use eight extracted features for classification, while deep learning approaches utilize the original music features for classification. All the experimental results are presented in Table 3. Machine learning models, such as Naive Bayes and decision trees, only achieved very low accuracy. This illustrates the difficulty early machine learning models face in achieving ideal results when dealing with large-scale data. Deep learning architectures, such as ResNet-50 and DenseNet-121, have demonstrated superior performance compared to conventional machine learning algorithms. This highlights the robust feature representation capability of deep learning models when trained on large-scale datasets. Furthermore, recent models for classifying musical genres have achieved satisfactory results but have not met expectations. The main reason for this is that these models require manual feature extraction, which limits the comprehensiveness of feature acquisition. This study proposed a combined deep-shallow model, in which deep learning models were used for feature extraction and shallow models such as SVM for classification. The experimental results demonstrated that the proposed approach achieved satisfactory results, outperforming existing models in musical genre classification.

Table 3. Experimental results

Models	Precision	F1-score	Specificity	Accuracy
Naive Bayes	0.5263	0.5556	0.5882	0.5050
KNN	0.9053	0.8269	0.7611	0.8218
Decision trees	0.6316	0.6250	0.6186	0.6436
Random forest	0.8737	0.8137	0.7615	0.8119
Logistic regression	0.7263	0.6832	0.6449	0.6832
CNN	0.8632	0.8159	0.7736	0.8168
SVM	0.7789	0.7475	0.7184	0.7525
ResNet-50	0.8947	0.7589	0.8213	0.8168
DenseNet-121	0.9263	0.7719	0.8421	0.8366
Foleis and Tavares [3]	0.8421	0.7921	0.7477	0.7921
Puppala et al. [5]	0.7895	0.7463	0.7075	0.7475
Singh and Biswas [19]	0.8526	0.8100	0.7714	0.8119
Yu et al. [20]	0.9263	0.8844	0.8462	0.8861
Liu et al. [21]	0.9368	0.8725	0.8165	0.8713
Proposed method	**0.9684**	**0.9154**	**0.8679**	**0.9158**

The bold font indicates the best performance in each test.

5. Discussion

In this study, we utilized a combined approach of DRAE and SVM to investigate the feasibility of musical genre classification. DRAE is an encoding architecture with excellent generalization ability and training efficiency. It has proven to be a powerful tool for automatic feature extraction. Simultaneously, an SVM was employed to classify and predict specific musical genres. The advantage of SVM is its powerful generalization ability, achieved by adopting the maximum margin hyperplane to separate nonlinear functions. In addition, SVM not only provides a better solution to nonlinear, high-dimensional, and limited training samples but also helps avoid overfitting problems, making it the best approach with high performance. A comparison with existing approaches indicates that the combined approach of DRAE and SVM utilizes the inherent information of musical genres in an optimal manner and achieves the best performance, with a classification accuracy as high as 0.9158. The novel approach suggests that when there is insufficient data, the DRAE is competent for feature extraction tasks, and the SVM is a competitive solution for classification in terms of accuracy and computational efficiency.

6. Conclusion

Musical genre classification is significant because it allows for the discovery of interesting genres within massive music datasets. This study investigated different music genres, analyzed existing approaches to musical genre classification, and proposed a hybrid machine learning approach that combines the DRAE and SVM classifiers. Eight conventional music features: audio RMS power, spectral centroid, spectral bandwidth, spectral roll-off, poly features, zero-crossing rate, tempogram, and Tonnetz, were manually extracted. Additionally, 20 features were obtained from the MFCC stage. The two main features were fused to create the initial input vectors for classification. All feature mappings were input into the DRAE for feature extraction, and the extracted features were then fed into the SVM for classification. This novel approach was inspired by the strong feature extraction capability, feature classification, and robustness of the SVM. Hybrid datasets, such as GTZAN, QQ music, and NetEase cloud music, were combined for training and testing. The experimental results indicated that the proposed method achieved a satisfactory F1-score of 0.9154 and an accuracy of 0.9158 in musical genre classification, outperforming existing approaches. This makes a significant contribution to the development of artificial intelligence in music. However, the proposed approach currently does not consider the items that when multi-genre on a single music. Therefore, even under the same setting, the presented approach only gives one musical genre on one input music data as lack of training data. Conceivably, the multi-genre on a single music is the classic multi-label learning area that is expected for us to solve in the future.

References

[1] R. Sarkar, N. Biswas, and S. Chakraborty, “Music genre classification using frequency domain features,’ in *Proceedings of 2018 5th International Conference on Emerging Applications of Information Technology (EAIT)*, Kolkata, India, 2018, pp. 1-4. https://doi.org/10.1109/EAIT.2018.8470441

[2] C. Weiss, F. Brand, and M. Muller, “Mid-level chord transition features for musical style analysis,” in *Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 341-345. https://doi.org/10.1109/ICASSP.2019.8682293

[3] J. H. Foleis and T. F. Tavares, “Texture selection for automatic music genre classification,” *Applied Soft Computing*, vol. 89, article no. 106127, 2020. https://doi.org/10.1016/j.asoc.2020.106127

[4] A. Vidwans, P. Verma, and P. Rao, “Classifying cultural music using melodic features,” in *Proceedings of 2020 International Conference on Signal Processing and Communications (SPCOM)*, Bangalore, India, 2020, pp. 1-5. https://doi.org/10.1109/SPCOM50965.2020.9179597

[5] L. K. Puppala, S. S. R. Muvva, S. R. Chinige, and P. S. Rajendran, “A novel music genre classification using convolutional neural network,” in *Proceedings of 2021 6th International Conference on Communication and Electronics Systems (ICCES)*, Coimbatre, India, 2021, pp. 1246-1249. https://doi.org/10.1109/ICCES51350.2021.9489022

[6] C. Chen and X. Steven, “Combined transfer and active learning for high accuracy music genre classification method,” in *Proceedings of 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, Nanchang, China, 2021, pp. 53-56. https://doi.org/10.1109/ICBAIE52039.2021.9390062

[7] J. L. Conceicao, R. de Freitas, B. Gadelha, J. G. Kienen, S. Anders, and B. Cavalcante, “Applying supervised learning techniques to Brazilian music genre classification,” in *Proceedings of 2020 XLVI Latin American Computing Conference (CLEI)*, Loja, Ecuador, 2020, pp. 102-107. https://doi.org/10.1109/CLEI52000.2020.00019

[8] N. Pelchat and C. M. Gelowitz, “Neural network music genre classification” *Canadian Journal of Electrical and Computer Engineering*, vol. 43, no. 3, pp. 170-173, 2020. https://doi.org/10.1109/CJECE.2020.2970144

[9] J. S. Luz, M. C. Oliveira, F. H. Araujo, and D. M. Magalhaes, “Ensemble of handcrafted and deep features for urban sound classification,” *Applied Acoustics*, vol. 175, article no. 107819, 2021. https://doi.org/10.1016/j.apacoust.2020.107819

[10] D. Taufik and N. Hanafiah, “AutoVAT: an automated visual acuity test using spoken digit recognition with MEL frequency cepstral coefficients and convolutional neural network,” *Procedia Computer Science*, vol. 179, pp. 458-467, 2021. https://doi.org/10.1016/j.procs.2021.01.029

[11] S. Wang, H. Wang, Q. Gao, and L. Hao, “Auto-encoder neural network based prediction of Texas poker opponent's behavior,” *Entertainment Computing*, vol. 40, article no. 100446, 2022. https://doi.org/10.1016/j.entcom.2021.100446

[12] M. Seo and K. Y. Lee, “A graph embedding technique for weighted graphs based on LSTM autoencoders,” *Journal of Information Processing Systems*, vol. 16, no. 6, pp. 1407-1423, 2020. https://doi.org/10.3745/JIPS.04.0197

[13] B. M. Aslahi-Shahri, R. Rahmani, M. Chizari, A. Maralani, M. Eslami, M. J. Golkar, and A. Ebrahimi, “A hybrid method consisting of GA and SVM for intrusion detection system,” *Neural Computing and Applications*, vol. 27, pp. 1669-1676, 2016. https://doi.org/10.1007/s00521-015-1964-2

[14] A. C. Enache and V. Sgarciu, “Anomaly intrusions detection based on support vector machines with an improved bat algorithm,” in *Proceedings of 2015 20th International Conference on Control Systems and Computer Science*, Bucharest, Romania, 2015, pp. 317-321. https://doi.org/10.1109/CSCS.2015.12

[15] Y. Chen and R. Zhang, “Default prediction of automobile credit based on support vector machine,” *Journal of Information Processing Systems*, vol. 17, no. 1, pp. 75-88, 2021. https://doi.org/10.3745/JIPS.04.0207

[16] H. Dai, J. Li, Y. Kuang, J. Liao, Q. Zhang, and Y. Kang, “Multiscale fuzzy entropy and PSO-SVM based fault diagnoses for airborne fuel pumps,” *Human-centric Computing and Information Sciences*, vol. 11, article no. 25, 2021. https://doi.org/10.22967/HCIS.2021.11.025

[17] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2015 [Online]. Available: https://arxiv.org/abs/1512.03385.

[18] Andrada, “GTZAN Dataset - Music Genre Classification,” 2022 [Online]. Available: https://www.kaggle.com/andradaolteanu/gtzan-dataset-music-genre-classification.

[19] Y. Singh and A. Biswas, "Robustness of musical features on deep learning models for music genre classification," *Expert Systems with Applications*, vol. 199, article no. 116879, 2022. https://doi.org/10.1016/j.eswa.2022.116879

[20] Y. Yu, S. Luo, S. Liu, H. Qiao, Y. Liu, and L. Feng, "Deep attention based music genre classification," *Neurocomputing*, vol. 372, pp. 84-91, 2020. https://doi.org/10.1016/j.neucom.2019.09.054

[21] C. Liu, L. Feng, G. Liu, H. Wang, and S. Liu, "Bottom-up broadcast neural network for music genre classification," *Multimedia Tools and Applications*, vol. 80, pp. 7313-7331, 2021. https://doi.org/10.1007/s11042-020-09643-6

Xue Han **https://orcid.org/0000-0002-6431-6026**

She received B.S. and M.S degrees in Arts from Harbin Normal University in 2006 and 2019 respectively. She is currently a lecture in Northeast Agricultural University. Her research interests include musical art and music artificial intelligence.

Wenzhuo Chen **https://orcid.org/0000-0003-0583-8761**

She is currently pursuing the B.S. degree in college of electrical & information, NEAU. She is the member of Intelligent Computing Training Center from Northeast Agricultural University. Her research interests include image processing and artificial intelligence.

Changjian Zhou **https://orcid.org/0000-0002-2094-6405**

He is as the supervisor in Intelligent Computing Training Center, and the director of Data and Computing Department from Northeast Agricultural University. His current research interests include music artificial intelligence and digital signal processing.