Hindawi
Mobile Information Systems
Volume 2022, Article ID 9203420, 9 pages
https://doi.org/10.1155/2022/9203420

Research Article

Artificial Neural Network for Folk Music Style Classification

Qinliang Ning[1] and Junyan Shi[2]

[1]*Music and Dance College of Hunan First Normal University, Changsha 410205, China*
[2]*Dongbang Culture University, Seoul 100-744, Republic of Korea*

Correspondence should be addressed to Junyan Shi; nqjy391025@hnfnu.edu.cn

Received 29 December 2021; Revised 26 January 2022; Accepted 18 February 2022; Published 21 April 2022

Academic Editor: Hasan Ali Khattak

Folk music style classification is of great significance. Traditional folk music style classification has difficulties in feature selection, and the existing folk music style methods based on deep learning also have shortcomings. In this paper, we use artificial neural networks to classify folk music styles and transform audio signals into a sound spectrum. In this paper, we use artificial neural networks to classify folk music styles and transform audio signals into a sound spectrum to avoid the problem of manually selecting features. Further, we combine the characteristics of the music signal and a variety of music data enhancement methods to enhance the music data. The proposed model can extract elements of the sound spectrum that are more closely associated with a certain music style category. Experimental results reveal that the proposed method achieves a high accuracy rate, which verifies the effectiveness of our model.

1. Introduction

Folk music is an aural communication that comprises and reflects human emotions. It is composed of rhythms, melody, harmony, or musical instruments fused according to a certain rule. Folk music is an art genre. Folk music genres are made up of diverse factors such as different rhythms, timbres, song patterns, and other qualities that combined compose folk music genres. Because of the fast growth and invention of the Internet and multimedia technologies in recent years, digital music has long been the primary form of music listening for people, fostering the rising demand for music appreciation. The folk music style is now one of the most often used categorization features for digital music database management and storage on most online music websites. The manual labeling procedure used in early music information retrieval is no longer efficient enough to meet management's demands to stay up with the ever-increasing volume of music data. Adding label information such as song lyrics, mood, instrument, and rhythm to each musical work as the analytical foundation in the recommendation system provided by the American PanDoRa Music Station Music Genome Project, for example, required a significant amount of human and time investment. Consequently, if we want to automate folk music style categorization, research into music style classification algorithms is crucial [1–5].

For this reason, since 2002, the folk music style categorization automation approach has been intensively researched as an important branch in the study of musical information retrieval as a beneficial practical tool for folk artists worldwide. In algorithm research for music style categorization, digital signal processing, random processes, and music theory are often employed. Numerous music properties, including genre-related qualities of a music signal, may be mathematically defined. A classifier is developed using machine learning approaches by first understanding the feature distribution characteristics of distinct styles. Finally, the classifier is given the characteristics of a particular audio signal segment as input, and the posterior probability is used to identify the style of the segment. The structure of the feature determines the upper limit of the classification algorithm's performance, and an effective representation technique may raise the classification result's accuracy to the maximum degree feasible. As a result, many scholars are researching the link between music signal engineering and its characteristics. Music signal

analysis, on the other hand, is complicated by two key difficulties. For one, music includes a plethora of complex, abstract information that is difficult to describe in artificially produced aspects such as rhythms, instruments, and chords. Musical signals contain a more intricate frequency composition and more timbre data than normal speech signals. As a consequence, traditional speech signal processing techniques cannot be applied directly to music signals, necessitating the creation of new algorithms tuned to their specific qualities [6–10].

Image, audio, and natural language processing have all benefited greatly from advances in deep learning in recent years. Deep neural networks are being used by an increasing number of academics in the hunt for a good feature expression for music signals. The algorithm's performance gains theoretical value while also replacing the previous manually derived features. Now that Spotify has the largest genuine streaming music service platform, deep learning has been used in its recommendation algorithm with great results. Deep learning-based music signal processing has the potential to accelerate the growth of music platforms while also improving the overall user experience, which has commercial and research benefits [11–15]. Deep learning developments in recent years have tremendously aided image, audio, and natural language processing. A growing number of academics are using deep neural networks in their search for a decent feature expression for music data. The algorithm's performance improves theoretically while simultaneously replacing previously obtained characteristics. Deep learning has been used in Spotify's recommendation system with excellent results now that it has the biggest real streaming music service platform. Deep learning-based music signal processing has the potential to accelerate the expansion of music platforms while also enhancing the overall user experience, which has both commercial and scientific implications [11–15]. In this paper, we use artificial neural networks to classify folk music styles and transform audio signals into a sound spectrum. In this paper, we use artificial neural networks to classify folk music styles and transform audio signals into a sound spectrum to avoid the problem of manually selecting features. Further, we combine the characteristics of the music signal and a variety of music data enhancement methods to enhance the music data.

2. Related Work

The music style categorization approach normally involves two steps of training and testing. In the training phase, initially, a mathematical model is built to characterize the distinguishing digital properties of musical genres. Then use preemphasis, Mel filtering, cepstrum boosting, and other approaches to extract the digital properties of the music file. Finally, the classifier is trained based on several styles of digital features and distribution characteristics. In the testing phase, digital feature extraction is conducted in the same manner as in the training phase. Use the classifier developed in the training phase to compute the retrieved digital characteristics and assess the style.

Early hidden Markov models and support vector machine models, K-nearest neighbor models, gradient boosting models, and extreme random tree models are examples of traditional machine learning music style models. Shao et al. [16] suggested utilizing the unsupervised clustering approach based on HMM to estimate similarity for music style recognition to characterize music data more thoroughly and create better identification results. The author uses Mel frequency cepstral coefficients and linear prediction cepstral coefficients and their associated first-order and second-order difference spectra to segment the inherent rhythmic structure of diverse musical genres. Then create an HMM for each song, and use agglomerative hierarchical clustering to complete the clustering operation [17]. Tempo and beat are regular parts of music's rhythm structure, demonstrating their relationship. The data has a distinct temporal sequence, which may be characterized using HMM. Using music attributes and the relative spectrum transformation-perceptual linear prediction feature [18], literature [19] constructs an active learning SVM model. We are working on improving the active learning approach for reducing uncertainty and creating a sample balancing standard. We may make the chosen samples more helpful by altering the sample balance while keeping their variety. With an accuracy rate of 81.2 percent, the active learning SVM model can recognize five distinct musical genres. The feature features are input into a KNN model, and tests are performed on the GTZAN dataset [20, 21] to see if the KNN model is successful. Reference [22] compares the Gradient Boosting and Extra Trees methods, demonstrating that both can retrieve multidimensional digital components.

Scholars have attempted to increase the accuracy of automated music style detection by using deep learning models. Deep neural networks' development has offered technological assistance for performing classification tasks on vast amounts of music data [23, 24]. Because of the large quantity of music data recordings available on the Internet at the time, the use of DNN to identify music types became the standard technique [25, 26]. However, DNN's flaws have steadily shown over time. Each piece of music, for example, has a distinct melody. If the melody is seen as a unique sequence of expression, then this characteristic is critical for distinguishing between various genres of music. The DNN model's training is mostly based on data from each frame and is incapable of adequately capturing the temporal dependency of the audio sequence. Although researchers have conducted some studies on the sequence representation of speech data [27, 28], they cannot be successfully applied to music data because the sequence features of voice and music are fundamentally different. One of the major research areas has been using recurrent neural networks to construct deep learning music style and recognition models. Literature [29] presented the notion of fusion segment features and showed the efficacy of segmented extraction of music features. It effectively brought long- and short-term memory networks into the area of music style identification. Literature [30] builds a successful LSTM and gated loop unit music style recognition model using two data formats for model training. The GRU model and the LSTM model are

contrasted. The GRU model demonstrated greater music style identification accuracy on the GTZAN data set. However, LSTM and GRU are difficult to convey music elements with high discrimination [31]; therefore, they are not suited for music style detection. In the realm of image recognition, a deep convolutional neural network model with weight sharing and excellent local perception properties is used. This allows you to understand the local and delicate musical style traits and the frequency rhythm of the music components in the music spectrum picture more effectively. The application of DCNN migration to the area of music style identification has also emerged as a current research focus.

3. Method

This section initially describes the music data processing procedure, including sound spectrum extraction, music data augmentation, and audio segmentation algorithms. Next, we introduce the proposed convolution structure, followed by an explanation of the general structure of the music categorization model based on convolutional neural networks. Finally, experiments are done to validate the model's usefulness.

3.1. Music Data Processing

3.1.1. Sound Spectrum Extraction. The Mel sound spectrum is used as the sound spectrum feature in this paper. The procedure of extracting the Mel sound spectrum consists mostly of the following three phases: (1) The acoustic signal is framed and windowed. (2) Using a short-time Fourier transform, get the sound spectrum. (3) After passing the sound spectrum through the Mel filter bank, the Mel sound spectrum is generated. To begin, the music's sound signal is transformed using the short-time Fourier transform. Then, apply the Mel scale to change the frequency on the amplitude spectrum. The amplitude is then transformed using the Mel filter. The outcome represents each frame's Mel spectrum, and the matching Mel spectrum is created by merging the analysis window's spectrum.

3.1.2. Music Data Enhancement. The number of parameters in deep neural networks much surpasses that of conventional machine learning models. If you want to ensure that the network's parameters have adequate values, you often want sufficient training data. Data augmentation technology enables the generation of more varied data from limited data, increasing the number of training samples and diversifying the data included in training samples, which is advantageous for network training. This paper directly improves the audio signal of music to guarantee that the increased audio adheres to the creative expression of music, does not impair the capacity of the original music to express itself artistically, and enriches the data set's content. In this paper, we investigate enhancing the audio quality of music data.

Audio overlay: the sound signal is created over time, and numerous sound-generating devices may vibrate concurrently. As with waves, sound waves propagate independently and are superimposed. The human auditory system can distinguish many sounds and musical components concurrently. We may superimpose several music samples of the same genre based on this sound property. Suppose the original signal of a piece of music is M_1 and the original signal of another piece of music in the same category is M_2M_2; the enhanced audio M_a after audio superposition can be obtained by

$$M_a = \alpha M_1 + (1 - \alpha)M_2, \tag{1}$$

where α is the scale factor and $\alpha \in (0, 1)$.

Audio speed control: the original music's pace is slightly increased or slowed to the original α times, the value of α is chosen at random from a uniform distribution of (0.9, 1.1), and the excess or shorter length is trimmed or filled. This method's rationale is that the tempo of songs of the same genre might vary somewhat in various playing environments. Slightly altering the song's playback speed has little effect on the overall style of the song, but it may lower the network's sensitivity to set rhythms and improve the model's flexibility.

Intensity adjustment: intensity adjustment refers to making a little alteration to the original music's loudness, boosting or reducing the original audio's loudness by $|\alpha|$ dB. The value of $|\alpha|$ is chosen from the nonzero integer range of (−10,10). Slightly adjusting the volume of the music will not affect the style of the song but will modify the absolute value of each point in the sound spectrum accordingly. It may also lessen the network's sensitivity to sound intensity and increase the model's capacity to adapt to changing audio loudness.

Pitch adjustment: consider that it is common to use different tones to play the same music in actual scenes. It changes the pitch will not bring significant changes to the original genre, so the pitch of the music can be adaptively adjusted. The pitch adjustment method adopted in this paper is to increase or decrease the audio frequency by $|\alpha|$ semitones, where the value of $|\alpha|$ is randomly selected in the nonzero integer interval of (−1,1). Given that it is typical in real scenarios to employ multiple tones to play the same song, because adjusting the pitch has no substantial impact on the original genre, the pitch of the music may be modified adaptively. The pitch adjustment technique used in this study raises or lowers the audio frequency by $|\alpha|$ semitones, with the value of $|\alpha|$ chosen at random from the nonzero integer range of (−1,1).

3.1.3. Audio Segmentation. The model in this section does not employ the whole sound spectrum created by the audio as its input but rather a shorter audio slice as the fundamental prediction unit. The following are the three goals of this process: (1) reduce the model's input size while increasing computation speed; (2) increase the size of training data, which is beneficial to model training; (3) the final prediction label of the audio sample is derived

from the prediction results of all audio slices, which aids in classification performance.

The rationale for audio segmentation stems from the concept that when individuals listen to music, they can discern between music genres without having to listen to the complete piece of audio. Using slices may also help to minimize the model's input size while allowing the model to concentrate more on gathering valuable local information. During the training phase, the model trains each audio slice individually once the audio has been sliced. Treating each audio slice as a distinct sample increases the quantity of the training data. In the prediction, the outcomes of all audio slices of music are combined, and the prediction's voting method is implemented.

When training the model, keep in mind that the same audio slice cannot appear in both the test and training sets at the same time. Because their audio signals are increasingly similar, using them for training and prediction at the same time will lead the model's predictive signs to become inaccurate. It may react more aggressively than the real scenario, resulting in an erroneous assessment of the model's performance. In this paper, we separate the data set used for training, prediction, or verification before slicing the audio to prevent this problem. This ensures that the identical audio slice does not exist in both the test set and the training set, nor in both the training set and the validation set at the same time. If the same audio slice appears in both the validation and training sets, the validation set's meaning is lost.

3.2. One-Dimensional Residual Gated Convolution Structure

3.2.1. Selection of Convolution Kernel. Probabilistic neural networks excel at detecting patterns in massive amounts of data. More abstract features may be formed in the network's deep layers by superimposing convolution kernels and executing repeated convolution operations. Convolutions are classified into two types: those that convolve in a single direction and those that convolve in several directions simultaneously. It is more critical to capture the translational invariance of data characteristics when utilizing one-dimensional convolution. This is usually how time changes when time-related data is analyzed. One-dimensional convolution is often used in sensor data analysis time series for signal data analysis within a set time period, such as audio signals.

The direction of translation decides whether one-dimensional convolution varies from two-dimensional convolution, and hence their calculation methods are not fundamentally different. When transformed to a sound spectrum, the original audio signal appears to be a single-channel grayscale image with a time sequence as

$$f_{w,h} = a\left(\sum_{i=0}^{k_w-1}\sum_{j=0}^{k_h-1} w_{i,j}x_{i+w,j+h} + b\right), \tag{2}$$

where a is the activation function, k_w is the width, k_h is the height, and b is the bias of the convolution.

3.2.2. Gated Linear Unit. Gated linear unit (GLU) [32] works as

$$Y = \mathrm{Conv}(X) \otimes \sigma(\mathrm{Conv}(X)), \tag{3}$$

where σ is the Sigmoid function.

The convolution-based gate control approach varies from the sophisticated gate control mechanism of the LSTM network. There is just a single input gate. This also speeds up the training of the network model based on the gated convolution unit.

3.2.3. Channel Attention. In this paper, we achieve the channel attention mechanism based on the SENets introduced in [33]. This network structure won the most recent ImageNet competition and performed very well on a variety of computer vision tasks. Researchers used it in conjunction with one-dimensional convolution to produce excellent results in music categorization tasks. The SENets network can pay attention to the connection between channels, and the model can automatically learn the relevance of various channel properties and realize the channel's attention mechanism.

It is primarily made up of two modules: Squeeze and Excitation. The Squeeze module corresponds to the global average pooling in Figure 1, compressing the temporal dimension to 1, and converting each channel to a real value. To finish the summary of time-domain features, calculate the statistical information related to each channel using the pooling procedure. The Excitation module's primary job is to understand the link between the channels and to realize the impact of the gating mechanism. The two fully linked layers learn the statistical information produced by the compression process for each channel and explicitly describe the correlation between the distinctive channels. The SE structure concludes with a reweighting operation that uses the weight generated by the Excitation module to determine the relevance of the relevant feature channel. The weighting of the features is done by multiplying the Sigmoid output value by the original input corresponding channel, and the channel-based attention mechanism is eventually realized.

3.2.4. One-Dimensional Residual Gated Convolution with Channel Attention. This section will present the fundamental convolution structure utilized in our proposed model, which incorporates the gated linear unit, residual connection, and channel attention mechanism mentioned before. First, consider the combination of gated convolution unit and residual connection. The residual gated convolution unit after adding the residual can be expressed as

$$Y = X + \mathrm{Conv}(X) \otimes \sigma(\mathrm{Conv}(X)). \tag{4}$$

The inclusion of residual structure in the gated convolution unit is intended to address more than only the issue of gradient disappearance. The most significant job is to transfer data over numerous channels. The network's multichannel transmission capacity has been strengthened by adding the residual connection to the gating unit.

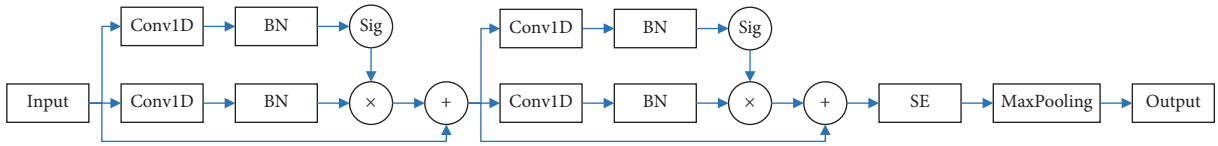

Figure 1: The structure of RGSE block.

Another reason why the SE structure is popular is that it is not intrusive to the network and is unnecessary to employ. The one-dimensional residual gated convolution structure paired with the channel attention mechanism utilized in the final model of this chapter is seen in Figure 1. In this paper, we refer to it as the RGSE block.

3.3. Structural Design of Convolutional Neural Network

3.3.1. Overall Network Structure. The RGSE block will serve as the foundation for the network module used in this paper. The model is divided into three sections: the RGSE block-stacking layer, the global pooling layer, and the input-to-output completely linked layer. Figure 2 depicts the structure.

3.3.2. RGSE Block Stacked Layer. This layer is mostly made up of a stack of numerous RGSE blocks, and the deep abstract elements of the sound spectrum are extracted by stacking convolution kernels. The RGSE block has 64, 128, or 256 convolution kernels, except for the one-dimensional convolution kernel used for channel number conversion in the first layer, which has a width of 5, and the one-dimensional convolution kernel in the RGSE block, which has a width of 3. In conjunction with the sound spectrum characteristics, this work employs a smaller convolution kernel for sound spectrum feature extraction, primarily for the following two reasons: (1) Deepen the network structure and improve the network's learning capabilities. Additional nonlinear layers ensure that the network can learn more complicated patterns in the input data. And, for a given input, the stacked tiny convolution kernel's receptive field is bigger than the receptive field of a single large-size convolution kernel. Stacking numerous tiny convolution kernels may produce the same result when utilizing one-dimensional convolution to handle sound spectrum data. Because stacking tiny convolution kernels is similar to constantly extending the perception range of the sound spectrum in the time domain, the number of network layers is greater when the perception range of the large-size convolution kernel in the time domain is the same. The network's capacity to discern probable patterns in the sound spectrum will improve as a result. (2) Simplify network parameters. The network parameters of numerous small-size convolution kernels are less than those of a large-size convolution kernel. This property is very significant for the gated convolution structure described in this paper. Because a gated convolution unit is equal to utilizing two parallel convolution kernels, the network parameters of the gated convolution unit will be twice as large as the network parameters of the conventional convolution unit at the same depth. If a large-size convolution kernel is utilized, the network's parameters will be raised further, which is detrimental to the network's training. As a result, to reduce the network parameters, a small-size convolution kernel must be used.

3.3.3. Global Pooling Feature Aggregation Layer. This layer is between the RGLU-SE block's fully linked and stacked layers. The higher layer's feature vector is subjected to one-dimensional global average pooling and one-dimensional global maximum pooling. The aggregate of pooling characteristics is then carried out.

The two forms of global pooling play distinct functions. In general, the impact of global maximum pooling is superior. Even though both maximum pooling and average pooling downsample the data, the maximum pooling procedure picks the greatest value for each unit in the feature map. Another feature selection may capture the most typical pattern in the feature map. The usage of numerous electronic musical instruments, for example, is the most evident element of electronic music. Assume the convolutional layer caught the pattern of a certain electronic musical instrument. In this situation, global maximum pooling might emphasize the pattern, making it easier to recognize electronic genres. However, maximal pooling has a drawback in that it only takes the essential region in each feature map. This will result in a feature map; even if just one region is associated with a certain music category, this feature map may significantly influence genre category prediction.

Global average pooling is somewhat bigger in terms of lowering the contribution of parameter dimensions. It is more favorable to comprehensive information transmission since it is more prone to downsampling the overall feature information. Furthermore, global average pooling considers each area, and the presence of one or two exceptional regions does not affect the final model's prediction outcomes. In this case, global average pooling allows the model to make better decisions. To effectively exploit the statistical information from the pooling layer, the model in this section combines global maximum pooling and global average pooling to generate a global pooling feature aggregation layer. The stacked layers of the RSE block provide a feature map subjected to the global average and global maximum pooling algorithms. The statistics of average and maximum pooling are calculated accordingly. Finally, the pooling procedure yielded just a one-dimensional output.

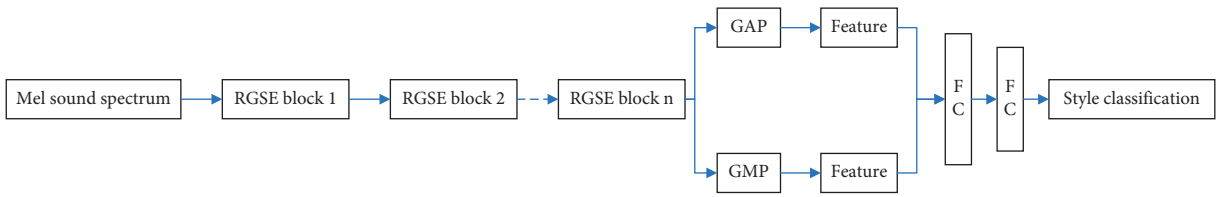

Figure 2: Network structure diagram.

3.3.4. Fully Connected Layer. The fully connected layer is at the model's conclusion, and the number of neurons in the output layer is equal to the number of music genres to be categorized. A batch standardization layer and a dropout layer were then added. Dropout will randomly disregard a fixed percentage of neurons in each training batch, decreasing the model's reliance on specific local characteristics and the danger of overfitting. The dropout probability of the fully connected layer is set at 0.2 in this chapter, and the activation function is ReLU.

4. Experiment and Discussion

4.1. Dataset. In this paper, the self-created ethnic music data collection comprises ten genres. Each style category has 100 audio samples, each sample lasts 30 seconds, and the sampling rate is 22050 Hz. The experiment separated the data set 8 : 1 : 1 into training, validation, and testing sets, ensuring balanced categories. The division is repeated ten times in this paper for tenfold cross-validation. In this paper, we use the classification accuracy (Acc) as the assessment metric for the proposed music style classification approach.

4.2. Evaluation of Convolutional Structure. We also investigate the impact of the RGSE composition structure on classification performance and contrast the convolution structure suggested in this paper with other types of convolution modules. For comparison, the following alternative structures were used in this experiment: (1) CNN uses regular one-dimensional convolution rather than gated convolution and lacks gated structure, residual connection, and channel-based attention mechanism. (2) GLU: unlike CNN, gated convolution is utilized instead of conventional convolution, and it lacks a residual link and a channel-based attention mechanism. (3) RCNN: it includes the same residual structure as RGSE but lacks a channel-based attention mechanism compared to CNN. (4) RGLU: it has the same residual structure as RGSE but lacks a channel-based attention mechanism compared to GLU. (5) RGSE: when compared to RGLU, the same channel attention method, namely, the convolution structure discussed in this paper, is added. Figure 3 depicts the end outcome.

When residual structure and gated convolution are coupled, more information is communicated in the network. The RGSE and SE structures increase classification accuracy. This demonstrates how paying attention to the various feature map channels may improve music categorization accuracy. The findings reported in this paper indicate that the RGSE convolution structure performs as intended.

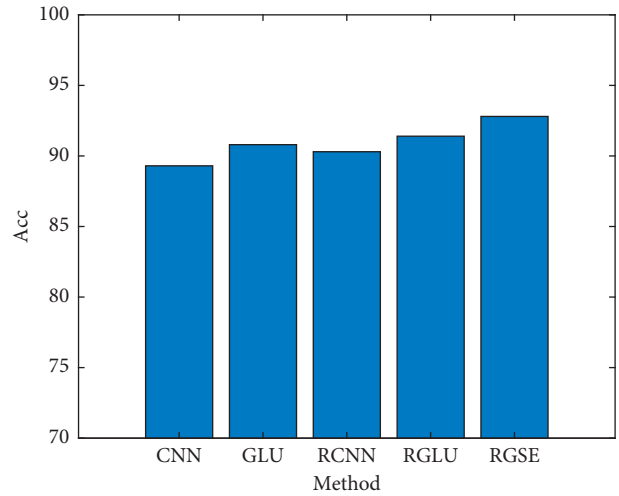

Figure 3: Classification accuracy of different convolution structures.

4.3. Evaluation of Global Pooling Feature Aggregation. This experiment analyzes the classification performance of different pooling features and their combinations in the global pooling feature aggregation layer. Three distinct pooling algorithms are tried on the feature map created by the RGSE block-stacking layer. Table 1 shows the results.

By combining the two global pooling features, the model's accuracy may improve. When using the global average pooling feature alone rather than the global maximum pooling feature, the entire statistical information of the spectroscopic feature map is more favorable to classification. We increase classification performance by providing the fully connected layer with additional statistical information about the features abstracted by the convolutional layer using two forms of global pooling features.

4.4. Evaluation of Audio Segmentation. In principle, adopting shorter audio slices as the primary training and prediction unit may increase the model's performance. Furthermore, the slice-based method allows the model to adapt to sounds of varying durations without further processing audio samples of varying lengths. Because the 50% overlap rate is typically used in comparable investigations, the experiment will begin by setting the overlap rate to 50%. Analyze the influence of slice duration on classification ability and then run the further analysis with varied overlap ratios. The best classification performance of the approach in this section is shown in Figure 4 for various slice sizes.

The model's classification performance is the poorest when the number of slices is 59 and the corresponding slice

Table 1: Comparison of different pooling methods.

Pooling method	Acc
GAP	91.7
GMP	90.9
GAP + GMP	92.8

time is 1 second, even though the number of samples and slice prediction results averaged at the prediction time has grown. However, the quantity of information included in one second of audio is insufficient for the model to complete music categorization, resulting in a reduction in the prediction performance of a single slice. Consequently, even if the prediction results of 59 audio sample slices were pooled, the accuracy could not be increased. The number of slices in an audio sample is 11 when the slice time is 5 seconds. To ensure that a single slice has enough information to offer a model for music categorization, the number of samples following the slice has been raised by more than tenfold. Because the slice length and number of slices are balanced, the model gets the greatest classification performance when the slice time is 5 seconds. When the number of slices is 1, the slice time is 30 seconds, implying that no audio is split. The classification accuracy is lower than when just a portion of the slicing is used. This finding suggests that selecting an appropriate slicing approach might improve the model's classification performance.

The related classification impact is stronger when the number of slices is 9, 11, and 14, and the corresponding slice lengths are 6 seconds, 5 seconds, and 4 seconds. The three slice durations are chosen to experiment with various overlap ratios, and the results are displayed in Figure 5.

The figures show that utilizing overlap in the three slice durations improves accuracy over not using overlap. This is due to the overlapping method's ability to keep continuous information on both sides of the segmentation point and its data augmentation impact of moving the audio signal. Furthermore, the findings show that just raising the overlap rate does not increase the model's classification performance. The amount of repeated data from various samples is too high, causing the model to oversample certain audio signals during training, affecting the final classification impact. In this experiment, the 50 percent overlap rate provided the greatest classification performance, demonstrating that utilizing an appropriate overlap rate is useful to increase model classification performance.

4.5. Evaluation of Music Data Enhancement. Compared to typical machine learning techniques, neural networks contain a much larger number of parameters. As a result, a substantial quantity of data support is required if the trained neural network model accurately matches the data distribution and has higher generalization performance. Although the use of slice-based training and prediction algorithms may help relieve the issue of data scarcity, data improvement is required to further extend the data amount. Furthermore, the improved audio contributes to the model's generalization capabilities. Figure 6 compares the increase in

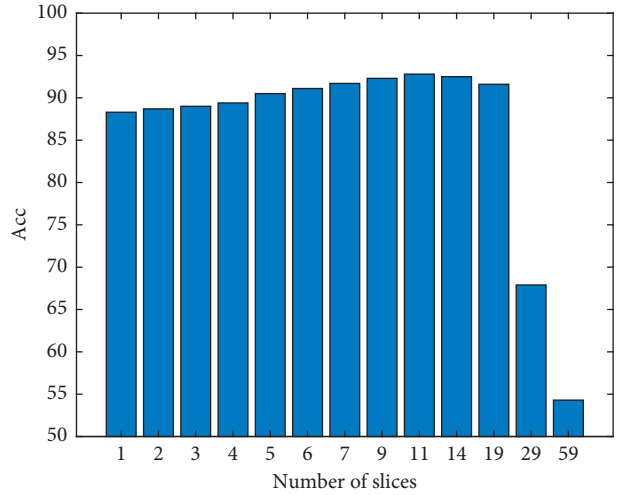

Figure 4: Classification accuracy under different number of slices.

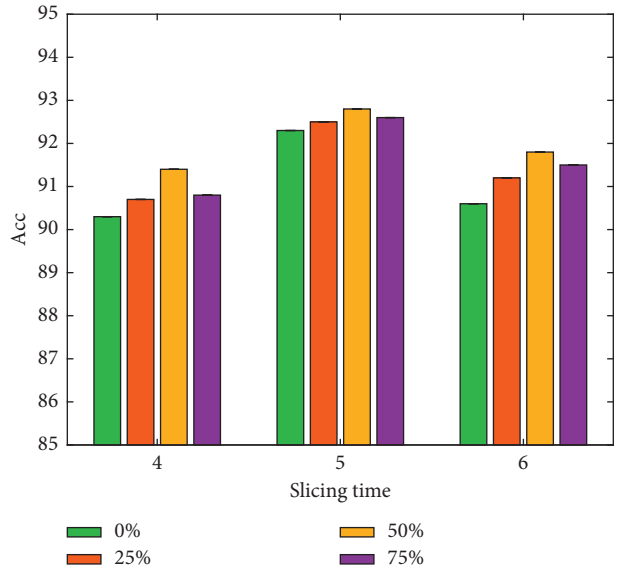

Figure 5: Classification accuracy of different audio segmentation overlap rates.

classification accuracy when utilizing the original data set with various audio enhancement approaches. Audio superimposition, audio speed adjustment, sound intensity adjustment, and pitch adjustment are all represented by the letters E1 to E4.

The image shows that employing audio overlay alone has the greatest impact since overlaying music signals of the same category into music offers a deeper form of representation. The second is pitch adjustment, which is a crucial aspect of music. This demonstrates that fine-tuning the pitch may increase the model's capacity to generalize. Combining four audio enhancement approaches may boost categorization accuracy even more. This experiment demonstrates the music enhancement strategy's effectiveness for music categorization in this study. It may give a richer music model for the data set while increasing the data set's size, which is beneficial to

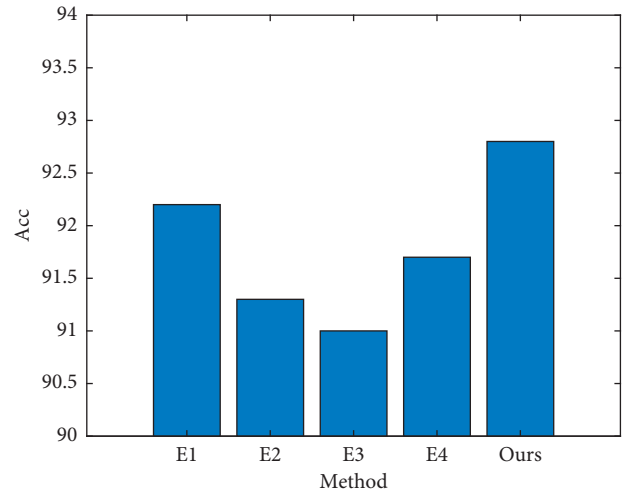

Figure 6: Accuracy of different music data enhancement methods.

network training and increases the model's music classification performance.

4.6. Evaluation of the Type of Sound Spectrum. This experiment will compare the Mel sound spectrum with the short-time Fourier sound spectrum, which is also commonly utilized in the area of music categorization. Table 2 compares the model's classification ability when various sound spectra are used and the experimental data are displayed.

As can be observed, using the Mel sound spectrum as a feature improves the classification performance of this technique. Because the perception of sound by the human ear's hearing system is not linear, the Mel sound spectrum exhibits more Mel scale transformation than the short-time Fourier sound spectrum. The Mel scale also employs a nonlinear frequency scale to detect equidistant pitch shifts, which corresponds to the properties of the human auditory system. This is why, in this experiment, the Mel sound spectrum outperforms the short-time Fourier sound spectrum.

4.7. Comparison with Other Methods. This section compares the method proposed in this paper with the music style classification baseline methods, as shown in Table 3.

In this paper, we propose that the RGSE network use a one-dimensional residual gated convolution structure paired with channel attention. Furthermore, the convolutional layer's GAP and GMP characteristics are pooled. The network can extract more important audio elements for the music category and obtain the best classification performance.

5. Conclusion

Aiming at the problem of the lack of music labeling data, in this paper, we combine the characteristics of music to directly enhance the original sound signal of music. The music's volume, pitch, and tempo are randomly modified by superimposing waveform signals of the same genre of music to generate music samples with deeper emotions. Our method offers a new form of convolutional neural network model by combining one-dimensional convolution, gating, residual connection, and attention. One-dimensional convolution takes advantage of the sound spectrum's temporal and frequency independence and the combination of residual structure and gating mechanism. It can help decrease network degradation and enhance the gate control unit's capacity to choose information flow. The channel-based attention allows the model to learn spectral properties from many channels. Simultaneously, the model aggregates global average pooling and global maximum pooling results, ensuring that the data is completely exploited [37].

Table 2: Classification performance of different sound spectra.

Sound spectrum	Acc
Mel sound spectrum	92.8
Short-time Fourier spectroscopy	90.3

Table 3: Comparison with other methods.

Method	Acc
IMFL [26]	84.5
MMDNN [34]	85.8
IMGC [35]	87.9
MGC [36]	91.2
Ours	92.8

Data Availability

The datasets used are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This project was supported by the Key Scientific Research Project of the Hunan Provincial Department of Education, Project no. 19a104.

References

[1] B. K. Baniya, D. Ghimire, and J. Lee, "Automatic music genre classification using timbral texture and rhythmic content features," in *Proceedings of the2015 17th International Conference on Advanced Communication Technology (ICACT)*, pp. 434–443, IEEE, PyeongChang, Korea (South), July 2015.

[2] D. C. Corrêa and F. A. Rodrigues, "A survey on symbolic data-based music genre classification," *Expert Systems with Applications*, vol. 60, pp. 190–210, 2016.

[3] G. Kour, N. Mehan, G. Kour, and N. Kakkar, "Music genre classification using MFCC, SVM and BPNN," *International Journal of Computer Application*, vol. 112, no. 6, pp. 12–14, 2015.

[4] B. K. Baniya and J. Lee, "Importance of audio feature reduction in automatic music genre classification," *Multimedia Tools and Applications*, vol. 75, no. 6, pp. 1–14, 2016.

[5] T. Makkonen, "Northern comfort: geographical scale, locality and the evolution of networks in the Finnish metal music genre," *Area*, vol. 47, no. 3, pp. 334–340, 2015.

[6] K. Zhang, "Music style classification algorithm based on music feature extraction and deep neural network," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 9298654, 7 pages, 2021.

[7] C. Weiss and M. Muller, "Tonal complexity features for style classification of classical music," in *Proceedings of the IEEE International Conference on Acoustics*, pp. 688–692, IEEE, Brisbane, QLD, Australia, April 2015.

[8] J. Pei, K. Zhong, J. Li, J. Xu, and X. Wang, "ECNN: evaluating a cluster-neural network model for city innovation capability," *Neural Computing & Applications*, pp. 1–13, 2021.

[9] D. Yang, K. Ji, and T. Tsai, "A deeper look at sheet music composer classification using self-supervised pretraining," *Applied Sciences*, vol. 11, no. 4, p. 1387, 2021.

[10] S. H. Mao and E. E. P. Myint, "Performance comparison of multi-class SVM classification for music cultural style tagging," *International Journal of Computer Theory and Engineering*, vol. 5, pp. 317–320, 2013.

[11] Q. G. Rafi, M. Noman, and M. Noman, "Comparative analysis of three improved deep learning architectures for music genre classification," *International Journal of Information Technology and Computer Science*, vol. 13, no. 2, pp. 1–14, 2021.

[12] S. Oramas, F. Barbieri, O. Nieto, and X. Serra, "Multimodal deep learning for music genre classification," *Transactions of the International Society for Music Information Retrieval*, vol. 1, no. 1, pp. 4–21, 2018.

[13] G. Song, Z. Wang, F. Han, S. Dinga, and X. Gu, "Music auto-tagging using scattering transform and convolutional neural network with self-attention," *Applied Soft Computing*, vol. 96, Article ID 106702, 2020.

[14] G. Gwardys and D. Grzywczak, "Deep image features in music information retrieval," *International Journal of Electronics & Telecommunications*, vol. 60, no. 4, 2014.

[15] J. Zhang, "Music feature extraction and classification algorithm based on deep learning," *Scientific Programming*, vol. 2021, Article ID 1651560, 9 pages, 2021.

[16] X. Shao, C. Xu, and M. S. Kankanhalli, "Unsupervised Classification of Music Genre Using Hidden Markov Model," in *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763)*, June 2004.

[17] J. Sheng and L. Xie, "Semi-supervised agglomerative hierarchical clustering based pairwise constraints," *Microcomputer & Its Applications*, vol. 31, no. 24, pp. 67–69, 2012.

[18] S. S. Upadhya, A. N. Cheeran, and J. H. Nirmal, "Thomson Multitaper MFCC and PLP voice features for early detection of Parkinson disease," *Biomedical Signal Processing and Control*, vol. 46, no. 9, pp. 293–301, 2018.

[19] X. Shao and L. Yao, "Music classification based on SVM active learning," *Computer Engineering and Applications*, vol. 52, no. 6, pp. 127–133, 2016.

[20] M. A. Ali and Z. A. Siddiqui, "Automatic music genres classification using machine learning," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 8, pp. 337–344, 2017.

[21] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[22] B. Murauer and G. Specht, "Detecting music genre using extreme gradient boosting," in *Proceedings of the Companion of the Web Conference 2018 on The Web Conference 2018. Palais des congrčs de Lyon: International World Wide Web Conferences Steering Committee*, pp. 1923–1927, Lyon, France, April 2018.

[23] I. H. Chung, T. N. Sainath, B. Ramabhadran et al., "Parallel deep neural network training for big data on blue gene/q," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 6, pp. 1703–1714, 2016.

[24] I. McLoughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, "Robust sound event classification using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 540–552, 2015.

[25] X. Yang, Q. Chen, S. Zhou, and X. Wang, "Deep belief networks for automatic music genre classification," in *Proceedings of the Florence Italy:Twelfth Annual Conference of the International Speech Communication Association*, vol. 92, no. 11, pp. 2433–2436, Florence, Italy, August 27-312011.

[26] S. Sigtia and S. Dixon, "Improved music feature learning with deep neural networks," in *Proceedings of the IEEE international conference on acoustics, speech and signal processing*, pp. 6959–6963, IEEE, Florence, Italy, May 2014.

[27] H. Wang, T. Lee, and C. C. Leung, "Unsupervised spoken term detection with acoustic segment model," in *Proceedings of the International Conference on Speech Database and Assessments*, pp. 106–111, IEEE, Hsinchu, Taiwan, October 2011.

[28] H. Wang, C. C. Leung, T. Lee, B. Ma, and H. Li, "An acoustic segment modeling approach to query-by-example spoken term detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5157–5160, IEEE, Kyoto, Japan, March 2012.

[29] J. Dai, S. Liang, W. Xue, and N. Chongjia, "Long short-term memory recurrent neural network-based segment features for music genre classification," in *Proceedings of the International Symposium on Chinese Spoken Language Processing*, pp. 1–5, Tianjin, China, 17-October 2016.

[30] J. Jakubik, "Evaluation of gated recurrent neural networks in music classification tasks," in *Proceedings of the International Conference on Information Systems Architecture and Technology*, pp. 27–37, Szklarska Poręba, Poland, September 2017.

[31] S. Ma, Q. Wuniri, and X. Li, "Deep learning with big data: state of the art and development," *CAAI Transactions on Intelligent Systems*, vol. 11, no. 6, pp. 728–742, 2016.

[32] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," 2016, https://arxiv.org/abs/1612.08083.

[33] J. Hu, L. Shen, S. Albanie, and G. Sun, "Squeeze-and-Excitation networks," in *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1, Salt Lake City, UT, USA, June 2018.

[34] P. Zhang, X. Zheng, W. Zhang et al., "A deep neural network for modeling music," in *Proceedings of the ACM on International Conference on Multimedia Retrieval*, pp. 379–386, Glasgow United Kingdom, April 2015.

[35] W. Zhang, W. Lei, X. Xu, and X. Xing, "Improved music genre classification with convolutional neural networks," in *Proceedings of the Conference of the international speech communication association*, pp. 3304–3308, San Francisco, USA, September 2016.

[36] H. Yang and W. Q. Zhang, "Music genre classification using duplicated convolutional layers in neural networks," in *Proceedings of the Conference of the international speech communication association*, pp. 3382–3386, Graz, Austria, September 2019.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2015, https://arxiv.org/abs/1512.03385.