Hindawi
Mobile Information Systems
Volume 2022, Article ID 2616220, 8 pages
https://doi.org/10.1155/2022/2616220

Research Article

Research on Music Emotional Expression Based on Reinforcement Learning and Multimodal Information

Lige Zhang[1] **and Zhen Tian**[2]

[1] *Moscow Academy of Art, Weinan Normal University, Shanxi, Weinan 714000, China*
[2] *School of Education Science, Weinan Normal University, Shanxi, Weinan 714000, China*

Correspondence should be addressed to Lige Zhang; zhanglige@wnu.edu.cn

Received 18 April 2022; Accepted 12 May 2022; Published 30 June 2022

Academic Editor: Chia-Huei Wu

With the continuous development of the research in the field of emotion analysis, music, as a common multimodal information carrier in people's daily life, often transmits emotion through lyrics and melody, so it has been gradually incorporated into the research category of emotion analysis. The fusion classification model based on CNN-LSTM proposed in this paper effectively improves the accuracy of emotional classification of audio and lyrics. At the same time, in view of the problem that the traditional decision-level fusion method ignores the correlation between modes and the limitations of dataset, this paper further improves the existing Thayer dimension emotional decision fusion method, takes the audio energy axis data as the main discrimination basis, and improves the accuracy of decision fusion classification. Based on the results of music emotion analysis, this paper further carries out the task of music generation. Based on the feature that there is often consistent emotional expression between music words and songs, a dual Seq2Seq framework based on reinforcement learning is constructed. By introducing the reward value of emotional consistency and content fidelity, the output melody has the same emotion with the input lyrics and good results are achieved. Compared with the ordinary Seq2Seq, the accuracy of our proposed model is improved by about 1.1%. This shows that the accuracy of the model can be effectively improved by using reinforcement learning.

1. Introduction

With the continuous development of technology, digital music has become the mainstream channel of mass music appreciation in the past two decades and has been widely spread through the Internet. Reports show that there are more than 600 million active users of online music in China alone. At the same time, with the popularity of short video applications such as TikTok, more and more users are becoming the main body of emotional expression using music. Under the background of such a huge user base and the continuous growth of online music library capacity, how to describe and calculate a piece of music efficiently and apply it to the fields of music intelligent recommendation and intelligent generation has become a problem of great research value.

Earlier, music emotion recognition mostly focused on analyzing the underlying physical characteristics of a piece of audio through advanced technology. For example, Beth et al. used the main feature MFCCs (Mel frequency cepstral coefficients) for speech recognition for music modeling, which verified the rationality of MFCCs in music recognition [1]. In the subsequent research, many researchers have performed a lot of work in music emotional classification based on audio, established the Hevner model [2], Thayer model and TWC model [3], and PAD model [4], tried to model from different dimensions to describe the emotion types that music can contain, and constantly pursued the rationality and accuracy of classification. At the same time, in terms of audio-based classification methods, most scholars have adopted common machine learning classification algorithms, such as k-nearest neighbors (KNN) [5], support vector machine (SVM) [6], and Bayesian method [7]. These statistical models based on mathematics are affected by the number of samples and classifications, which is easy to lead to unsatisfactory results. Moreover, the traditional machine

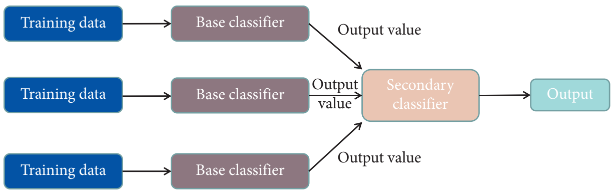

Figure 1: Basic framework of stacking.

learning methods also face the problems of high cost of audio feature extraction and inability to deal with large samples. Based on this, researchers began to try using deep learning methods such as Gaussian mixture model (GMM) [8] and convolutional neural network (CNN) [9] to classify music emotion. At the same time, in order to solve the problem of low accuracy of traditional deep learning methods, researchers began to seek some composite methods to classify music emotion more accurately. For example, Tang et al. improved the efficiency and performance of music classification tasks by combining in-depth learning with extensive learning [10]. At the same time, some researchers have carried out some work on emotional classification of music based on lyrics, most of which are based on machine learning methods of natural language processing. Chen et al. used the vector space model to judge the pressure level of lyrics in music, thus pioneering the single mode of lyrics to complete the recognition of music emotion [11]. The original electroencephalogram (EEG) signals of people receiving music information are directly applied to convolutional neural network and long-term memory network (CNN-LSTM), and a higher classification accuracy is obtained [12]. With the continuous development of related research on music emotional classification and the improvement of computer processing ability, single-modal music emotional classification research can no longer meet the performance requirements. Therefore, more and more researchers have begun to explore multimodal music emotional classification methods and confirm that the application of related methods reflects certain efficiency and use value. A common method of multimodal analysis is to classify music emotions by combining the two modes of lyrics and audio. Yang extracted both audio and text features, including MFCC. A multimodal fusion model based on V-A emotional space is proposed. After classifying music emotions in both audio and text dimensions, the results are linearly overlaid to analyze music emotions. However, there are some limitations in this kind of research. That is, the emotional connection between lyrics and songs is separated, and the consistency between them is ignored [13]. In terms of music generation, Jiang et al. proposed a deep reinforcement learning algorithm for online accompaniment generation, which makes it have the potential of real-time man-machine duet improvisation, and proved its music generation ability due to the baseline method through the method of subjective preference evaluation [14].

To sum up, this paper builds a multimodal interactive network based on the existing multimodal music emotional analysis methods, taking into account the emotional consistency of various modes, so as to obtain more ideal music emotional analysis results. At the same time, this paper also tries to produce music based on reinforcement learning, which can express specific emotions and is consistent with the emotional expression of lyrics.

2. Musical Emotional Classification Based on Multimodal Information

2.1. Multimodal Affective Classification Model

2.1.1. Stacking Method Introduction. Based on the analysis of music, it is divided into audio and lyric text. Secondary fusion of the audio features of songs and the text features of lyrics can generate the emotional fusion features of songs, which can more accurately judge the emotions of music and better reflect the emotional information and emotional tendencies contained. Among the multimodal music emotional classification methods combining audio and lyrics, there are three main types of multimodal fusion: data-level fusion, feature-level fusion, and decision-level fusion. Decision-level fusion is the most advanced one. Bagging [15] and Boosting [16] are both integrated learning methods in the field of machine learning. They are used to fuse many weak classifiers into strong classifiers. They are not fused for multimodal data among themselves but can be used effectively. However, this architecture cannot be used in deep learning methods and lacks generalization ability, the classifiers available in it have poor ability to process music feature data, and their practical applicability is low.

To solve the problems above, we make improvements to the basic feature-level and decision-level methods to avoid the loss of emotional information and solve the problem of feature relevance. Stacking [6] is a model integration technique that combines the outputs of multiple models to produce a new model, which combines multiple models to improve the results of machine learning. This method allows better classification performance than a single model and is not essentially a multimodal fusion approach. The core idea is to train the original sample features with different base classifiers, combine the results of the training classification labels, represent the new dataset sample features, then input the secondary classifier for learning training, and output the integrated classification results. The basic framework of stacking is shown in Figure 1.

The stacking method usually uses different base classifiers to produce heterogeneity of feature output values, and because of the smoothness of the integrated model, the integration performance of the integrated model is usually better than that of any base classifier model. Moreover, the integrated model can focus on the model that performs well and not trust the model that performs poorly. Therefore, the stacking method is very effective for integration of very different basic models. Therefore, this paper uses the stacking method to fuse the multimodal features in music.

Audio and lyric text in music are two different modes of feature representation. Although there is some emotional and semantic association, there is considerable heterogeneity in the digital representation of the feature data. The classification output by feature fusion directly is not good.

Stacking, on the other hand, just needs a variety of differentiated feature classification models as its base classifier. Traditional stacking inputs are a feature set of the same mode, resulting in different classifier output results. We attempt to use the audio and lyric classification model as a base classifier as a whole, and the stacking method can also be used to classify the output.

In the process of dividing the dataset for training output and converting it into the training set again, the stacking integration method is prone to overfitting if the entire training set of the training model is used to predict the label of the training set in turn. The model in the diagram uses the 5-fold (5-fold cross validation) method to solve the overfitting problem in the stacking process. The diagram of the model is shown in Figure 2.

2.1.2. Model Description

(1) Dataset Processing. The number of dataset samples used in this experiment is 2000, which is divided into the training set and dataset according to 8 : 2 ratio. Based on the original training set, the training set is further divided into five parts by 5-fold cross validation, as follows:

$$\begin{pmatrix} tr\_1 \\ tr\_2 \\ tr\_3 \\ tr\_4 \\ tr\_5 \end{pmatrix} (\text{test}). \tag{1}$$

(2) Base Classifier Training. The model contains two base classifiers: one is based on CNN-LSTM audio classification model (M1) and the other is based on text classification model (M2). First, for the audio classifier (M1), four (1280) parts of the training set are trained by the 5-fold cross validation in the Figure 2 and the remaining one (320) is predicted as follows:

$$\begin{pmatrix} tr\_1 \\ tr\_2 \\ tr\_3 \\ tr\_4 \end{pmatrix} \xrightarrow{\text{test}} (tr\_5) \xrightarrow{\text{predict}} (pr\_1). \tag{2}$$

The training model is also trained on the original entire test set:

$$\begin{pmatrix} tr\_1 \\ tr\_2 \\ tr\_3 \\ tr\_4 \end{pmatrix} \xrightarrow{\text{test}} (\text{test}) \xrightarrow{\text{predict}} (te\_1). \tag{3}$$

Five new *pr* and *te* copies were obtained by performing the above operations five times. Five copies of *pr* were connected, and the final P1 size was 1600. The average of five *te* samples was calculated, and the size of T1 was 400. We have the following:

$$(pr\_1) + (pr_2) + (pr_3) + (pr_4) + (pr_5) \longrightarrow \begin{pmatrix} pr_1 \\ pr_2 \\ pr_3 \\ pr_4 \\ pr_5 \end{pmatrix} \longrightarrow (P1),$$
$$\frac{((te\_1) + (te\_2) + (te\_3) + (te\_4) + (te\_5)}{5} \longrightarrow (T1). \tag{4}$$

For the text classification model (M2), P2, T2, P, and T (the same size as the original dataset, with 2 columns of features and all of the classification label information) are obtained by doing the same operations as P1 and T1 to connect the training set and the test set of the secondary classifier as follows:

$$\begin{aligned} (P1) + (P2) &= \begin{pmatrix} P1 & P2 \end{pmatrix}, \\ &= (P), \\ (T1) + (T2) &= \begin{pmatrix} T1 & T2 \end{pmatrix}, \\ &= (T). \end{aligned} \tag{5}$$

(3) Secondary Classifier Training. After the training of the base classifier, a new training set P and a new test set *T* are generated to train the secondary classifier. Data of different modes are fused into tags, eliminating the heterogeneity of feature data. Selection of subclassifiers is generally a logical regression method. Complex neural network classification models have been used in the process of output and combination of basic classifier features, so the selection of subclassifiers does not need to be too complex. This model uses Softmax layer [9] as the training method of the secondary classifier to make the final multimodal subclassification. The training description is as follows:

$$(P) \xrightarrow{\text{test}} (T) \xrightarrow{\text{predict}} (\text{output}). \tag{6}$$

In summary, the stacking framework integrates classification algorithms of different modes and synthesizes the ability of different classifier algorithms to extract features from different angles, so as to complement each other and optimize the results. This paper uses the stacking-based multimodal integration method to solve the problem of heterogeneity of different modal features. The integration results are more stable and accurate than those of the feature-level fusion method, and the model program is simple to implement. It does not need to adjust the single-modal classification model built before, nor does it need too many adjusting parameters to effectively combat overfitting.

2.2. Musical Emotional Classification Based on Audio Features

2.2.1. Extraction of Audio Features

(1) Audio preprocessing

① *Separation of Vocals from Melodies*. In order to explore a finer granularity and to easily reflect the

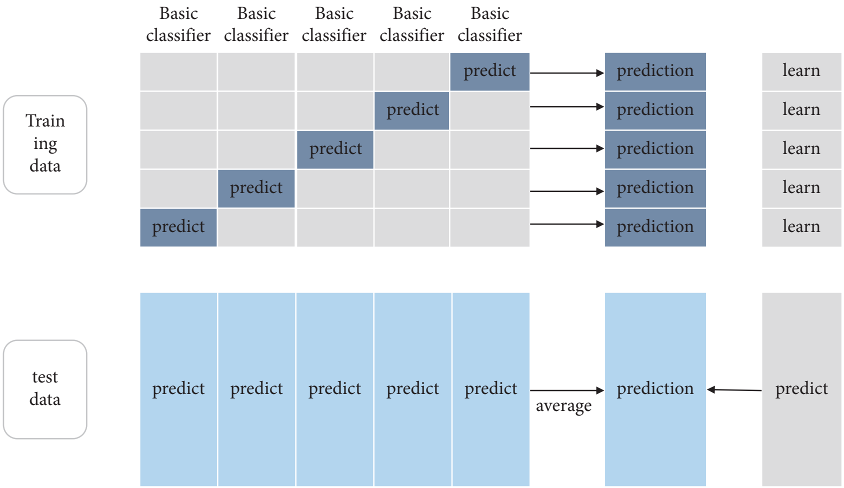

Figure 2: Stacking multimodal integration method.

relationship between melody and lyrics, it is necessary to separate vocals from melodies in a song. We use the open source program Spleeter (the address is https://github.com/deezer/spleeter) as a tool for separating vocals from melodies in music. Spleeter can evaluate each sound source by using a U-shaped network and use it as a soft mask, eventually separating the vocals from the piano, guitar, and other instruments. Based on the above methods, in the experimental process of this paper, the actual entire music is divided into four levels to build a dataset and the results of the classifier output are voted on. The first is an average segmentation of 30 s, the second is a fine-grained 15 s sentence-level segmentation, and the other two methods obtain pure vocal and background sound segments extracted by audio-processing tools. To improve the performance of emotional classification systems, experiments were performed to compare the classification results of different preprocessing methods.

② *Fine-Grained Slicing.* Long duration will result in too-large feature dimension and slow training speed, and the classifier is prone to overfitting. Moreover, music audio may show different emotional tendencies in different time periods, and direct emotional classification as a whole may lead to partial immersion. In order to synthesize audio emotional information and improve the speed of classification, this paper makes a fine-grained segmentation of the real music dataset and outputs the emotional results through voting decision-making, which can effectively improve the accuracy of music emotional classification.

(2) Selection of audio features

In the song audio, a variety of feature parameters can be extracted in time and frequency domains, among which the spectral feature combines the time-domain and frequency-domain characteristics and can well represent emotional information. A common solution is to generate spectrograms using short-time Fourier transform (STFT). However, music is different from voice data, the audio-converted spectrogram contains complex information, it is difficult to train, and the image feature representation is often related to the image resolution and other features. In practical research, there are some limitations in classifying the output of spectral graphs as audio emotional feature representations. Based on the actual needs of music emotion extraction, related research usually extracts both low-level and high-level descriptive features of audio. Low-level descriptors (LLDs) are low-level features that are designed by hand and are generally calculated from a single frame of audio. High-level statistics functions (HSFs) are features based on LLDs, such as mean and maximum. They are the feature representations for the multiframe audio. The specific feature selection is shown in Table 1.

2.2.2. Audio Emotional Classification Model Based on CNN-LSTM.

Based on the theme of music emotional classification, audio emotional classification often needs to combine both spectral and temporal characteristics. Because of the existence of convolutional and pooling structure, the convolutional neural network has a strong ability to synthesize information and extract features from two-dimensional data and can further compress features, while the cyclic neural network has the ability to process serialized feature data. This paper builds a fused affective classification model based on CNN-LSTM, which can be used to classify and output emotional feature data.

The convolutional and pooling layers in the CNN play a role in feature extraction and feature selection. A set of eigenvectors can be output using part of the structure of the CNN and input to SVM and LSTM as new features. Based on the fusion classification model of CNN-LSTM, audio features are input into the network. Spectrographic features extract features and select features through convolutional and pooling layers in the CNN. A set of serialized feature vectors are output, which are input into the LSTM network as new features and an attention mechanism is added for output. LLDs are combined with HSFs by statistical methods and then reduced by DNN. Finally, the output eigenvectors of the two network structures are lengthwise stitched into audio fusion features and classified with the Softmax layer to get the classification results. The model network architecture is shown in Figure 3. The model consists of two main parts: spectrogram + CNN-LSTM and LLDs + DNN. The model combines CNN's strong ability to extract comprehensive features from two-dimensional data and RNN's ability to extract context from time-series data. It further extracts spectral features from image and time-series features. Considering the insufficiency of classification ability of individual spectral features, LLD features are combined to compensate for emotional information representation in the network to improve classification performance.

2.3. Musical Emotional Classification Based on Lyrics

2.3.1. Extraction of Lyric Features

(1) TF-IDF feature extraction

Term frequency-inverse document frequency (TF-IDF) is a feature extraction method for weight representation based on the frequency of word occurrences in a file. TF-IDF can use probability statistics to calculate word occurrences, assess the proportion of word items in the document, determine the importance of the word, and use it to represent the emotional polarity of the lyric text. The more times an emotional representative word appears in a lyric text, the more important it is in the emotional classification evaluation. By combining all the frequency information, the emotional tendency of the whole lyric text can be evaluated. However, there are some drawbacks. It treats words in the document as separate features, ignoring the connections between words and the whole article.

(2) Feature extraction of the chi-square test

The chi-square test feature extraction method is derived from the chi-square test statistical method in mathematical statistics to describe the correlation between two random variables. In the expression of lyric text in songs, for a specific emotional type of lyrics, there are often a large number of compact descriptors. Statistical processing of these special emotional words can improve the performance of lyric text classification.

2.3.2. Lyrics Affective Classification Model Based on CNN-LSTM. The model is divided into two parts: word vector + CNN-LSTM and word frequency weight + DNN. First, the convolution neural network is used to extract the multiword vector features of the input text, and the extracted features are integrated into the input of the LSTM neural network to output a new set of word vector feature representations. Then, the bag model vectors extracted by TF-IDF or the chi-square test are extracted by DNN. The features of the two categories are stitched together as a fusion representation of lyric text and then classified and output by Softmax to get the result of text emotional classification.

Table 1: Different levels of audio description features.

Categories	Characteristics
LLDs	MFCC, zero crossing rate, spectral centroid, spectral bandwidth, chromaticity characteristics
HSFs	Maximum, mean, variance

Similar to the audio affective classification model, the two single-modal classification models are composed of input layer, CNN layer, LSTM layer, attention mechanism layer, DNN layer, and output layer and the main difference is the input layer. The input of the audio classification model is spectrogram and LLD features, while the input of the text classification model is word vector and word frequency weight vector. The framework of lyric emotional classification model is shown in Figure 4.

2.4. Multimodal Music Emotional Classification Experiment

2.4.1. Selection of the Dataset. The dataset used in the multimodal music emotional classification experiment is consistent with that used in the single-modal experiment by contrast. In the experiment of multimodal music emotion classification, the emotion tags (four), audio and lyrics text files are extracted by using the download tool. Emotional tags were also angry, happy, relaxed, and sad, with 500 of each affective list totaling 2,000, which are shown in Table 2. Random partitioning of the entire dataset is done, 80% of which is a training set and 20% of which is a test set.

2.4.2. Comparative Experiment. We chose some of the more mainstream classification methods as our comparative experiment to verify the validity and superiority of the proposed framework in music emotional classification. Results are shown in Table 3.

The results of the experiment show that the model is more effective than some of the mainstream research methods found in recent years. The single-modal research methods in the literature have achieved certain classification performance, in which the accuracy of lyrics classification is overall higher than that of audio classification, but there are limitations. The results of multimodal fusion experiments show that the feature-level and decision-level fusion methods can combine the emotional information of different modes and have better performance than single-modal fusion methods.

3. Music Generation Based on Enhanced Learning

In order to better apply the key element of music emotion, after the music emotion analysis, this paper attempts to study

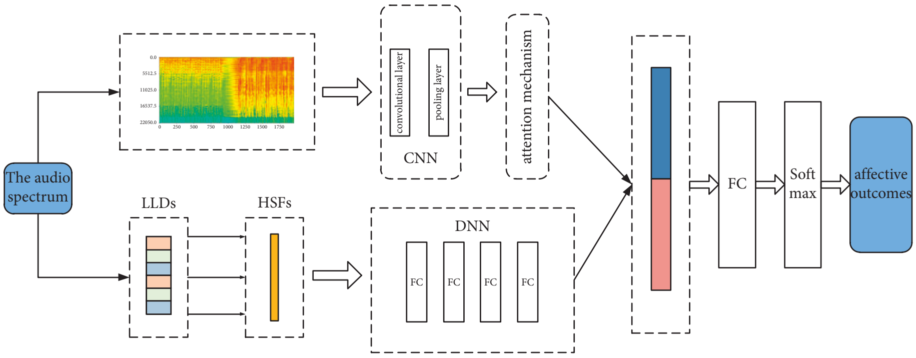

Figure 3: Audio emotional classification model based on CNN-LSTM.

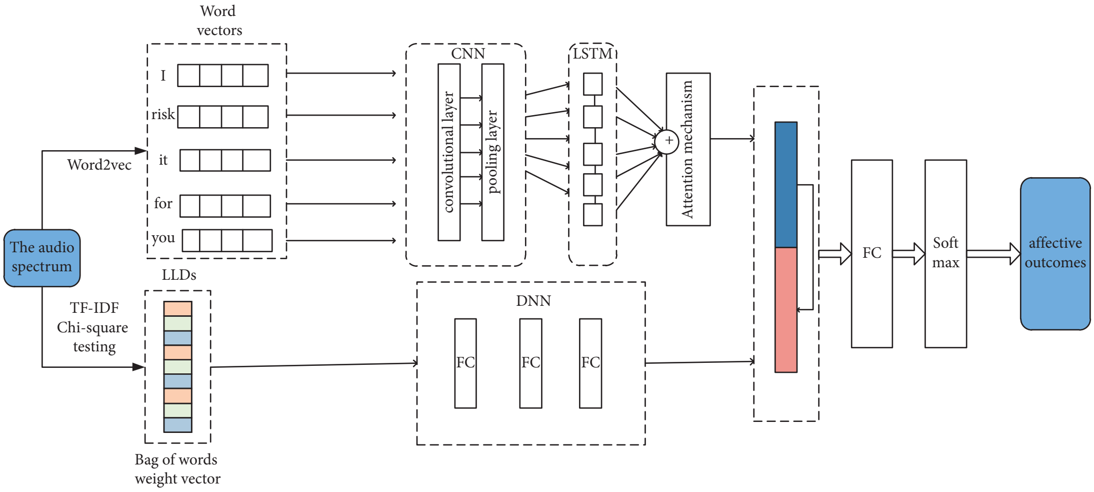

Figure 4: Lyric affective classification model based on CNN-LSTM.

music generation from the perspective of music emotion, to achieve the generation of music melodies of specified emotion types. We first improve the representation of existing MIDI word granularity datasets and then construct a dual Seq2Seq model based on reinforcement learning, which can constrain whether the emotions of input text and output audio are at the same level by adding emotional consistency and content fidelity constraints. The latter can make the model more stable for conversion. The end result is an input text and an output audio with the same emotions as the text.

3.1. Dataset Selection and Preprocessing. We selected a Chinese pop music dataset created by Lee et al. in 2019 [20]. They collected 1000 pieces of Chinese pop music and converted the melodies into the MIDI format to facilitate subsequent research. At the same time, word granularity alignment was done between MIDI and lyrics. These label data are very high-quality resources that can be used for music generation based on MIDI. On the basis of these data, the quality of music generation can be greatly improved, and each word can be given a specific note. Although the original dataset contains 1000 pieces of Chinese pop music, some of the music only has the label of the pitch note or only the label of the duration of the note. There are also songs where the number of note lengths per sentence does not match the number of notes and there is a loss. We filtered the dataset and ultimately kept 840 of them as the object of analysis.

Table 2: Dataset of multimodal music emotional classification.

Dataset	Angry	Happy	Relaxed	Sad
Training set	400	400	400	400
Test set	100	100	100	100

3.2. Seq2Seq Music Generation Model Based on Enhanced Learning. In music generation tasks, we also need to ensure that the emotions of lyrics and melodies are within the same range. For this purpose, we build a Seq2Seq music generation model based on reinforcement learning, which is composed of two Seq2Seqs, where one end is melody and the other end is lyrics. Its Reward is mainly composed of two parts; one part is emotional consistency Reward abbreviated as Rc, and the other part is content fidelity Reward abbreviated as Rs. The process of this framework is mainly divided into two steps: pretraining of multimodal affective model and training of dual Seq2Seq. The framework of Seq2Seq music generation model is shown in Figure 5.

In the task of music generation, we use the reinforcement learning method to restrict lyrics and melodies emotionally by setting reward function Reward, so that the emotions of

Table 3: Comparative experimental results between different models.

Time	Modal	Classification model	Accuracy
2019	Audio	LLDs + SVM [17]	0.571
2019	Lyric	Word2vec + LSTM [18]	0.693
2021	Multimodal	Random forest [19]	0.778
Model of this paper	Multimodal	Multifeature CNN-LSTM network + stacking fusion	0.782

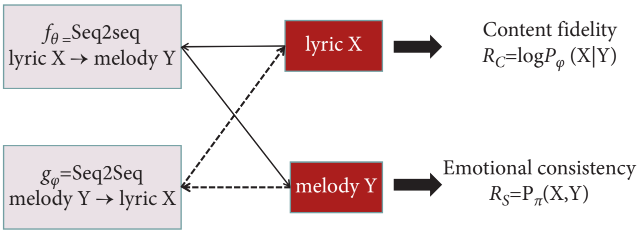

Figure 5: Overall framework of the Seq2Seq music generation model based on emotion consistency restriction.

Table 4: Comparison experiments on Remi representation.

	Lyrics granularity-1	Lyrics granularity-3
Seq2seq	0.562	0.184
Dual seq2Seq based on reinforcement learning	0.579	0.195

Table 5: The data scale of different datasets.

	Lyrics granularity-1	Lyrics granularity-3
Training set	10000	3248
Test set	957	458

the generated lyrics or melodies can be emotionally consistent with those of the input side. We improve the multimodal emotional classification model proposed in Section 3 as a discriminator of emotional consistency. In addition, for a better training model, we use the dual mirror training method to train two Seq2Seq models and introduce another content fidelity Reward to measure the training quality of two Seq2Seq models.

3.3. Experimental Results and Analysis. We use ACC to evaluate the performance of our model on two different partitioned datasets. We have strict requirements for accuracy, requiring that the Remi-digitized sequence generated be exactly the same as the digitized sequence of the original melody to be accurate and that the music we generate has the same chord, note, duration, and other information as the original melody. Because Lee uses two sequences of notes and duration to represent the MIDI audio on this original dataset, which cannot be directly compared, we use Seq2Seq, which is the same parameter in this paper, to perform comparison experiments on Remi representation. The experimental results are shown in Table 4, and the accuracy (ACC) is the evaluation index. The data scale of different datasets under this division is shown in Table 5.

4. Conclusion

Music contains abundant human emotional information. Studies about music emotional classification can help to organize and retrieve a large amount of music data. Music contains two modes of emotional information: audio and lyric text. By building a multimodal music emotional classification system, the classification performance can be effectively improved.

This paper chooses the Thayer emotional model as the basis of music emotional classification and divides music into four categories: anger, joy, relaxation, and sadness. To solve the problem of long duration and complex composition of real music, this paper presents a fine-grained segmentation preprocessing method and extracts pure background sound fragments through vocal separation to optimize the sample set to improve classification performance. In view of the limitations of single feature and the limitations of single network classification method, this paper presents a single-modal emotional classification model based on CNN-LSTM, which achieves the best classification results, with the audio classification accuracy of 68% and the text classification accuracy of 74%. The heterogeneity between different modal data is also a big challenge for the classification system of multimodal fusion. This paper presents a multimodal integrated learning method based on stacking architecture, which achieves the best multimodal classification performance with a classification accuracy of 78%, which is much better than that of single-modal classification.

At the same time, we designed a set of dual Seq2Seq models based on reinforcement learning. With this model, we can effectively control the emotion between lyrics and melodies to achieve the goal of music generation based on emotional guidance. The model adds reward values for emotional consistency and reward values for content fidelity to achieve the same emotional type between the output melody and the input lyrics.

In the future, we will mainly try to carry out from the following aspects. First, we will continue to improve the constructed music dataset, such as labeling some other music structures (such as prelude and bridge segment). At the same time, the classification method proposed in this paper should be improved and the optimization parameters of the network model need to be adjusted to improve the classification accuracy.

Data Availability

The labeled dataset used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This study was supported by Weinan Normal University.

References

[1] B. Logan, “Mel frequency cepstral coefficients for music modeling,” in *Proceedings of the International Symposium on Music Information Retrieval*, Plymouth, Massachusetts, USA, October 2000.

[2] G. Liu and Z. Tan, “Research on multi-modal music emotional classification based on audio and lyirc,” in *Proceedings of the 2020 IEEE 4th information technology, networking, electronic and automation control conference (ITNEC)*, pp. 2331–2335, IEEE, Chongqing, Cambridge, MA, USA, June 2020.

[3] A. Tellegen, D. Watson, and L. A. Clark, “On the dimensional and hierarchical structure of affect,” *Psychological Science*, vol. 10, no. 4, pp. 297–303, 1999.

[4] N. Song, H. Yang, and P. Wu, “A Gesture-To-Emotional Speech Conversion by Combining Gesture Recognition and Facial Expression recognition,” in *Proceedings of the 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pp. 1–6, IEEE, Beijing, China, May 2018.

[5] M. Sudarma and I. G. Harsemadi, “Design and analysis system of KNN and ID3 algorithm for music classification based on mood feature extraction,” *International Journal of Electrical and Computer Engineering*, vol. 7, no. 1, p. 486, 2017.

[6] C. Chen and Q. Li, “A Multimodal Music Emotion Classification Method Based on Multifeature Combined Network classifier,” *Mathematical Problems in Engineering*, vol. 2020, Article ID 4606027, 11 pages, 2020.

[7] F. H. Rachman, R. Sarno, and C. Fatichah, “Music emotion classification based on lyrics-audio using corpus based emotion,” *International Journal of Electrical and Computer Engineering*, vol. 8, no. 3, pp. 2088–8708, Article ID 1720, 2018.

[8] Y. R. Pandeya, B. Bhattarai, and J. Lee, “Deep-learning-based multimodal emotion classification for music videos,” *Sensors*, vol. 21, no. 14, p. 4927, 2021.

[9] Y. R. Pandeya and J. Lee, “Deep learning-based late fusion of multimodal information for emotion classification of music video,” *Multimedia Tools and Applications*, vol. 80, no. 2, pp. 2887–2905, 2021.

[10] H. Tang and N. Chen, “Combining CNN and broad learning for music classification,” *IEICE - Transactions on Info and Systems*, vol. E103.D, no. 3, pp. 695–701, 2020.

[11] R. H. Chen, Z. L. Xu, and Z. X. Zhang, “Content based music emotion analysis and recognition,” in *Proceedings of the 2006 International Workshop on Computer Music and Audio Technology*, November 2006, Article ID 68275.

[12] S. Sheykhivand, Z. Mousavi, T. Y. Rezaii, and A Farzamnia, “Recognizing emotions evoked by music using CNN-LSTM networks on EEG signals,” *IEEE Access*, vol. 8, Article ID 139345, 2020.

[13] D. Yang and W. S. Lee, “Music Emotion Identification from lyrics,” in *Proceedings of the 2009 11th IEEE International Symposium on Multimedia*, pp. 624–629, IEEE, San Diego, CA, USA, December 2009.

[14] N. Jiang, S. Jin, Z. Duan, and C. Zhang, “RL-duet: online music accompaniment generation using deep reinforcement learning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 710–718, NY, USA, February 2020.

[15] N. P. Urmi, N. U. Ahmed, M. H. R. Sifat, S. Islam, and A. S. M. M. Jameel, “BanglaMusicMooD: A Music Mood Classifier from Bangla Music Lyrics,” in *Proceedings of the International Conference on Mobile Computing and Sustainable Informatics*, pp. 673–681, Kirtipur, Nepal, January 2020.

[16] Y. R. Pandeya, B. Bhattarai, and J. Lee, “Music video emotion classification using slow-fast audio-video network and unsupervised feature representation,” *Scientific Reports*, vol. 11, no. 1, Article ID 19834, 2021.

[17] Y.-S. Seo and J.-H. Huh, “Automatic emotion-based music classification for supporting intelligent IoT applications,” *Electronics*, vol. 8, no. 2, p. 164, 2019.

[18] H. Piliang and R. Kusumaningrum, “Music Emotion Classification Based on Indonesian Song Lyrics Using Recurrent Neural Network,” in *Proceedings of the 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS)*, pp. 1–4, IEEE, Semarang, Indonesia, October 2019.

[19] S. Hizlisoy, S. Yildirim, and Z. Tufekci, “Music emotion recognition using convolutional long short term memory deep neural networks,” *Engineering Science and Technology, an International Journal*, vol. 24, no. 3, pp. 760–767, 2021.

[20] H. P. Lee, J. S. Fang, and W. Y. Ma, “iComposer: An Automatic Songwriting System for Chinese Popular music,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 84–88, Minneapolis, USA, June 2019.