A Multimodal Music Feature Extraction and Classification Method Based on Deep Learning

Feiyan Yang

School of Education and Sports, Guangzhou Sontan Polytechnic College, Guangzhou 511370, China, 13051292596@163.com

Corresponding author: Feiyan Yang, 13051292596@163.com

Abstract. This article conducted data application and task performance optimization analysis of multiple distribution categories on a multimodal dataset. The impact construction of the model in different periods was achieved by implementing a long short-term classification plan for the training data step size. On the basis of ensuring the efficient operation of the classification model, this article compares and analyzes the performance of other machine models. This article compares the accuracy of the LSTM model and support vector machine with the random forest model and finds that the performance of the LSTM model is more accurate than the other two models. The multimodal model features have shown an improvement in accuracy analysis after emphasizing the fusion of different states. With the assistance of computer verification of the recall rate, a critical investigation and verification were conducted on the fusion state data distribution of multiple models. Improving the classification scheme of the model greatly enhances the accuracy of related technologies in practical applications.

Keywords: Deep Learning; Long-Term Memory Network; Multimodal Data Classification; Feature Fusion; Performance Optimization
DOI: https://doi.org/10.14733/cadaps.2025.S3.1-14

1 INTRODUCTION

At present, research on using multimodal information for music emotion classification is still relatively weak. Some studies have proposed some simple multimodal fusion methods that comprehensively utilize the information of lyrics and audio modalities to classify music [1]. As the main component modality of music, most researchers are currently dedicated to using commonly used machine learning methods to study the role of audio information in emotion classification. However, the potential for effectively utilizing multimodal information in the field of music emotion classification has not been fully explored [2]. Only a small number of researchers have focused on the impact of lyrics on music emotion classification. Most of them use similar feature methods to represent song information, and the results show that lyrics can also play a significant role in music emotion classification. The experimental results show that the accuracy of emotion classification can be improved to a certain extent by using the information from multiple modes compared with using only one mode [3]. At present, research on multimodal problems has been widely applied in various fields such as event detection. These multimodal music emotion classification studies are still in their early stages, and there is still a huge semantic gap between low-level audio and text features and high-level emotional expression in music. The heterogeneity of feature spaces between different modalities makes it difficult for researchers to fully explore the correlations between modalities for sentiment classification. The existence of these problems makes multimodal music sentiment classification still a huge challenge [4]. These methods have improved the effectiveness of music classification to some extent, but in some special cases, feature extraction is still difficult. The core of the BCRSN model lies in its ability to capture local spatial features from the two-dimensional time-frequency representation (spectrogram) of music audio signals, and also learn the evolution laws of these features in the time dimension through recursive mechanisms, namely sequence information. And integrate these classification results through a weighting mechanism to achieve comprehensive and accurate prediction of music emotions [5]. This method cleverly transforms the originally complex regression prediction task into a series of parallel binary classification problems, each focusing on a specific aspect of musical emotion. This strategy not only significantly reduces computational complexity and shortens training time but also enhances the model's ability to capture different emotional dimensions through task segmentation [6]. In order to further optimize the efficiency and accuracy of the model, some scholars have designed an innovative Weighted Mixed Binary Representation (WHBR) method. This series of achievements fully demonstrates the enormous potential of combining multimodal music feature extraction with advanced deep-learning models.

The automatic classification of music, as a unique branch in the field of speech signal processing, its complexity is not only reflected in the general challenges of speech signal recognition technology [7]. Some scholars have delved into learning methods based on multimodal music feature extraction, aiming to comprehensively analyze the principles, methods, and technological frontiers of music classification. Although the rapid development of speech recognition technology in recent years has brought many new methods and tools to music classification, the diversity and uncertainty of music remain the main obstacles hindering its large-scale practical application. Moreover, the rich emotions, diverse styles, and high artistry inherent in music itself make music classification a highly challenging research topic. By systematically analyzing the limitations of traditional feature analysis algorithms, we propose a series of improvement strategies, particularly by combining multimodal feature extraction techniques with feature analysis methods [8]. By simulating the performance of music classification in different scenarios, we found that the integrated method based on multimodal music feature extraction not only achieved significant improvement in classification accuracy. In the experiment, we used various different feature sets and combined them with advanced classifiers for combination testing. The music in the music library is in MIDI format and includes styles from different genres such as jazz, classical, pop, and country. In order to better train music and obtain its features, single-track music will be chosen here. These music files are stored on computer hard drives, which can be either single hard drives or hard drive arrays depending on the size of the music library. Although the music in the music library is in MIDI format, some of it still does not meet the required criteria, so it still needs to be selected and processed. Because multi-track music contains many different instruments and chords, it is still difficult to generate such complex music using computers at the current level [9]. The main function of music features is to extract features that represent music genres, including pitch, loudness, and timbre. Data vectorization is mainly used to represent the features of music in the form of vectors, and to obtain more accurate music content, it is necessary to divide the time series into smaller equidistant intervals [10]. The expression of musical features in each genre is different, and the final features of the music will be learned through the input of feature vectors. The main purpose of model training here is to adjust the values of parameters so that they can reach the optimal level.

Although music feature recognition has received widespread attention, it is still a concept that is difficult to define clearly. This is because they are artificially categorized as fashionable and controversial. The fuzzy definition of music feature recognition may result in multiple classifications for a single song [11]. The traditional methods for music feature recognition usually include a preprocessing stage, striving to utilize as much information as possible. At the same time, the lack of consensus among human annotators on the classification of music makes it difficult for artificial classification to be universally recognized. On the other hand, it is not feasible for listeners to manually annotate a large amount of music data. Therefore, this vague definition and lack of real and reliable data make music feature recognition a daunting task. The machine learning stage reflects the selectivity of features, and its performance largely depends on the features extracted in the preprocessing stage. However, extracting effective music features is quite difficult, which requires people to fully grasp the relevant knowledge of music topology [12]. As is well known, deep learning algorithms have achieved great success in computer vision and natural language processing, and more and more deep learning algorithms are being applied to music feature recognition tasks. In addition, manually extracted features lack universality, and manually designed features for one task often perform poorly in another task.

One advantage of deep learning models is that they can automatically extract high-level features related to specific tasks from raw or processed data. Some scholars have proposed a feature fusion algorithm based on multi-level local feature encoding for music feature recognition tasks. This method uses local representation to capture different levels of local information and learn their dependencies. In addition, using feature-level fusion to learn the implicit relationships between high-order features extracted by the feature encoding network can capture the information exchange relationships between features in the early stages. We also considered the complementarity of scattering conversion features and transfer features relative to typical features, enriching the diversity of features and enabling the model to learn more comprehensive feature representations. Feature extraction does not require a deep understanding of music signals to extract effective music features, and the extracted features are transferable [13]. Some audio genre recognition methods based on ensemble learning involve fusion at the decision level during model integration, which may overlook the early information exchange of features. Neglecting the difference between music feature recognition and image recognition. Because music feature recognition is composed of highly diverse intermediate characteristics with different levels of abstraction. Secondly, some existing methods only consider a single audio feature and ignore the complementarity between diverse features, which cannot provide sufficient discriminative information for music feature recognition. Overall, existing music feature recognition models still have certain limitations. Firstly, current methods based on deep learning models focus on global feature learning, making decisions on features at the same level as typical image classification.

2 RELATED WORK

In the vast world of music creation, the main melody is not only the soul carrier of emotions but also an important embodiment of the content and ideological depth of the work. This method aims to go beyond a single-dimensional analytical framework and achieve a more comprehensive and accurate analysis of music melodies by integrating multiple music feature patterns. Through meticulous feature engineering, the most representative features were selected from the original music melody signals, and an optimized subset of melody features was constructed. In terms of detection principles, Li et al. [14] fully considered the fluidity, repeatability, and variability of melodies and automatically captured and learned the spatiotemporal distribution patterns of these feature points through the deep learning mechanism of TCN. In addition, in order to further improve the accuracy and efficiency of melody feature point detection, we innovatively introduce a Time Convolutional Network (TCN) and combine the advantages of multimodal feature extraction to propose a new melody signal feature point detection algorithm. A good foundation has been laid for the automated generation of different genres of music. The use of computers to automatically generate different styles of music is an important and popular field in music information retrieval and music production. More and more researchers are investing in the field of automated music generation, and currently, automated music generation has been applied by many composers. Prior to this, most researchers used deep learning networks for music genre classification and recognition. Now, more researchers are starting to use deep learning for music generation, so it is meaningful to study music generation of different genres. Due to the excellent performance of LSTM networks in dealing with long-time series problems, Liao and Gui [15] applied their relevant deep learning knowledge to design and implement an algorithm model that can generate multiple genres of music, based on their understanding of LSTM networks. In the preprocessing process, techniques such as track separation and track stitching are used to obtain music features, including timbre, pitch, and loudness. At the same time, the music data was quantified and the input and output data formats were designed. In the field of music genre style recognition and generation based on the LSTM network, Maba [16] proposed an innovative network design, which focuses on reimagining the network architecture and implementing a mechanism for all music genre subnets to share the decoding layer. This design not only significantly reduces the number of parameters that the model needs to learn, but also greatly improves learning efficiency, bringing important technological breakthroughs to the field of music style recognition and generation. In the parameter selection of the network, the influence of the number of hidden layers and the number of neurons in each layer on the experimental results was compared through experimental methods, and the optimal network parameters were finally found. At the same time, how other parameters were selected was also introduced. Each music genre subnet analyzes music from different genres, achieving the function of multitasking simultaneously. Through experiments, the optimal dropout coefficient was determined, and a matrix containing music features was generated using test data. A script was written to convert the music matrix into playable music. By analyzing the frequency spectrum and spectrogram of the generated music sequence and the original music sequence, it is demonstrated that the network has good performance in generating music of different genres. At the same time, a comparison was made between using the RBM method and the method proposed in this paper to generate music effects, demonstrating the superiority of this method.

Traditional methods are often limited by excessive reliance on data labels and limited one-dimensional analysis (such as frequency domain or physical features). Through sparse decomposition techniques, independent instrument components can be extracted from complex mixed instrument music data. These features not only intuitively and accurately express the composition of each instrument, but also contain deep information about the emotional and dynamic changes of music. They are like keys, unlocking the hidden emotional codes and creative intentions in musical works, and providing powerful tools for music analysis, emotional recognition, and composition assistance. This process not only reduces information redundancy but also enhances the interpretability and separability of features.

In order to significantly improve the comprehensiveness of theater music data collection and the accuracy of genre recognition, and to solve the problem of traditional algorithms being unable to effectively distinguish complex music genres due to the limitation of a single feature dimension. Panda et al. [17] utilized the characteristics of DBN to deeply explore and integrate multimodal music features, achieving precise recognition of music genres. It designs a music data collection scheme based on the Internet of Things, which utilizes a sensor network distributed throughout the theatre (such as sound sensors, environmental sensors, etc.) to capture music signals and surrounding environmental information in real-time. Pei and Wang [18] innovatively combined Internet of Things (IoT) technology with Deep Trust Networks (DBN), introduced a multimodal music feature extraction strategy and constructed an efficient and accurate music data collection and genre recognition system. In the data preprocessing stage, we not only focus on the spectral characteristics of audio signals, but also introduce multidimensional features such as time-domain features, timbre texture, rhythm patterns, dynamic changes, and emotional labels to form a multimodal music feature set. These features collectively depict the comprehensive style of music, laying a solid foundation for deep learning and precise classification of DBN. DBN, as a powerful deep learning model, can effectively capture high-order abstract features in data by stacking multiple restricted Boltzmann machines (RBMs) and top-level supervised learning layers. Provide rich and diverse data sources for subsequent multimodal feature extraction, such as audience reactions and stage lighting changes. Xiang [19] reduced the risk of model overfitting by randomly discarding some neurons in the network. Momentum optimization accelerates the convergence speed during the training process and improves the learning efficiency of the model. After carefully designed network architecture and parameter adjustments, a DBN model suitable for multimodal music feature input has been successfully constructed. This model can effectively capture and distinguish the feature differences of different genres in complex music datasets, achieving high-precision genre recognition.

Currently, DL models, notably Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), are extensively employed in music classification tasks. Notably, the LSTM network stands out as a potent tool for music feature extraction and classification due to its proficiency in handling sequential data. LSTM's ability to capture long-term dependencies in music sequences facilitates the extraction of more discriminatory features.

3 MULTI-MODAL MUSIC FEATURE EXTRACTION AND CLASSIFICATION METHOD

3.1 Feature Extraction

Audio features are fundamental and pivotal elements in music analysis. This study utilizes MFCC as its audio feature, a feature widely employed in automatic speech recognition and music information retrieval, effectively capturing the frequency spectrum characteristics of audio. Extracting MFCC features entails thorough audio signal preprocessing and a series of transformation steps. Initially, the audio signal is segmented into shorter frames. Windowing minimizes spectrum leakage, while pre-emphasis balances the spectrum and enhances high-frequency resolution. The audio signal is then converted from the time domain to the frequency domain using the fast Fourier transform, resulting in a complex array indicating the power of each frequency component. The power spectrum is obtained by squaring the modules of these complex numbers, representing the intensity of each frequency component. This power spectrum is subsequently processed through Mel filter banks, a collection of evenly spaced triangular filters on the Mel scale, mimicking the nonlinear frequency perception of the human auditory system. Each filter's output signifies the energy within its corresponding Mel scale band.

For the mth Mel filter, the centre frequency f_m can be calculated by the following formula:

$$f_m = \frac{f_{\max}}{In\ 1+F\,/\,Mel\ f_{\max}} \cdot In\left(1+\frac{m}{F \cdot Mel\ f_{\max}}\right) \tag{1}$$

Where $f_{\max}$ is Nyquist frequency, F is the number of filter banks and $Mel\ f$ is the conversion function from frequency to Mel scale:

$$Mel\ f\ = 2595\log_{10}\ 1+f\,/\,700 \tag{2}$$

Each vector element corresponds to a dictionary word and signifies its occurrence frequency in the lyrics. This representation disregards word order and grammatical structure, focusing solely on word frequency, thereby achieving effective digitization of lyrics.

The TF-IDF method is a more detailed text feature extraction method. It not only considers the frequency of words in a single lyric text (i.e., word frequency TF) but also considers the distribution of words in the whole data set (i.e., inverse document frequency IDF). By giving each word a weight, the TF-IDF method can reflect the importance of words in lyrics. Specifically, the more frequently a word appears in the lyrics and the less frequently it appears in the whole data set, the more it is considered to represent the uniqueness and importance of this lyric.

Music itself is composed of audio waveforms and lyrics without direct visual elements, but the audio features of music can indirectly reflect its style, genre, and overall emotional atmosphere. Audio feature extraction is an important step in understanding music works, as it reveals the intrinsic properties of music by analyzing audio signals. This study specifically focuses on the unique value of music feature extraction and explores the use of deep learning techniques. Like text or audio frames, but in music feature extraction, we may combine other networks, such as CNN, to capture key features in the audio. In the process of music feature extraction, we can use deep learning models to automatically learn and recognize advanced features in audio signals, such as spectral content, rhythm patterns, harmony structures, and dynamic changes. These features are crucial for distinguishing music works of different styles and genres. By training a specialized deep learning model, we can convert each piece of music audio into a high-dimensional feature vector. This feature vector not only contains the core information of the audio signal, such as frequency distribution, rhythm intensity, and timbre characteristics, but also succinctly represents key attributes related to music style, genre, and emotion. These feature vectors can then be used for various music analysis tasks, such as genre classification, emotion recognition, recommendation systems, etc., providing a new way of music understanding based on audio features.

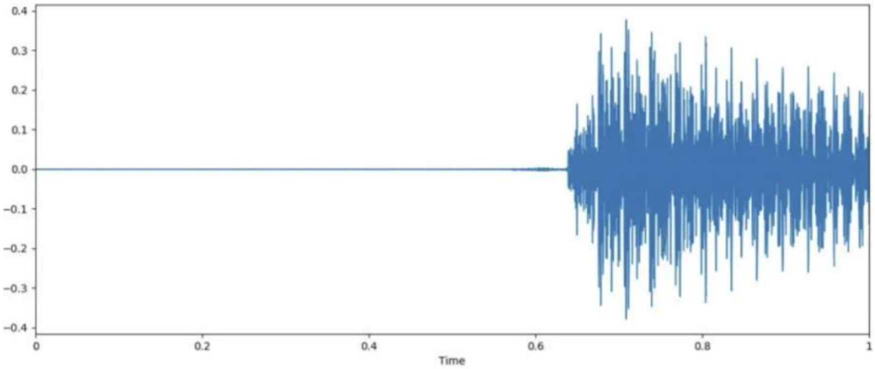

Figure 1: Music feature extraction.

Figure 1 shows the feature extraction process of music audio. After being processed by the LSTM model, it is converted into an information feature vector. This feature vector can be used for various music classification tasks, such as genre recognition, sentiment analysis, etc., providing us with a new way of music understanding based on visual information. These feature vectors not only contain the core features of audio signals, such as spectral features, temporal features, and harmonic structures but also represent key attributes related to music style, genre, emotion, etc. in a concise form. These feature vectors play an important role in subsequent music analysis tasks, such as genre recognition, sentiment analysis, song recommendation, etc., providing us with a new way of understanding music based on audio features.

3.2 LSTM Model Construction

The forget gate's role is to determine which information from the cell state should be discarded. Mathematically, this is expressed as:

$$f_t = \sigma \ W_f \cdot \left[h_{t-1}, x_t\right] + b_f \tag{3}$$

$$i_t = \sigma \ W_i \cdot \left[h_{t-1}, x_t\right] + b_i \tag{4}$$

Then, a tanh function creates a new candidate vector:

$$\tilde{C}_t = \tanh \ W_c \cdot \left[h_{t-1}, x_t\right] + b_c \tag{5}$$

$$C_t = f_t \cdot C_{t-1} + i_t + \tilde{C}_t \tag{6}$$

The LSTM layer excels at capturing long-term dependencies in sequence data, making it particularly adept at processing music due to its inherent time series characteristics. The network receives a multi-modal feature vector as input. After processing through several LSTM layers, music classification results are output through a fully connected layer. This layer employs the Tanh activation function to generate the probability distribution across each category. The overall architecture of the model is depicted in Figure 2.

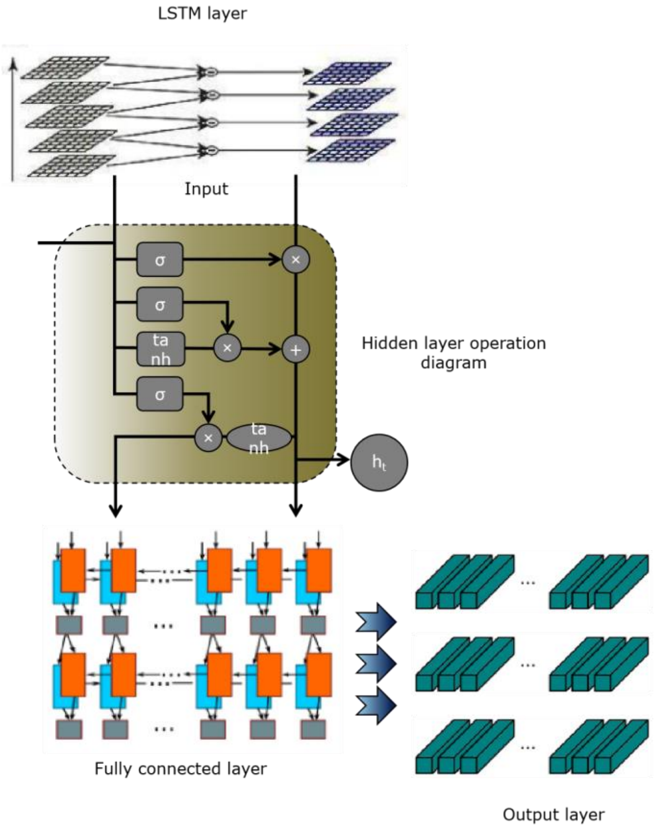

Figure 2: Overall structure of the LSTM model.

In the LSTM model, the cell state C_t runs through the whole time step, and only a small linear combination will change with time. The output gate determines which part of the cell state will be output to the next hidden state:

$$o_t = \sigma\ W_o \cdot \left[h_{t-1}, x_t\right] + b_o \tag{7}$$

Then, the cell state gets a vector with a value between -1 and 1 through a tanh function:

$$\tilde{h}_t = \tanh\ C_t \tag{8}$$

Finally, the combination of the output of the output gate and the cell state after tanh determines the hidden state:

$$h_t = o_t \cdot h_t \tag{9}$$

By controlling the inflow, outflow and retention of information through these gates, LSTM can effectively learn long-term dependencies in sequence data.

In the process of model training, this study uses the cross entropy loss function as the optimization goal and updates the network parameters through the backpropagation algorithm. Table 1 shows the key parameter settings of the LSTM model in training:

Parameter Name	*Parameter Value*	*Description*
Learning Rate	0.001	The specific value of the learning rate used for the Adam optimizer
Optimizer	Adam	The optimization algorithm used for model training
Batch Size	64	The number of samples input during each training iteration, balancing computational efficiency and memory usage
Epochs	100	The number of times the entire dataset is traversed to ensure sufficient model learning
Dropout Rate	0.5	The dropout ratio is set to prevent overfitting, indicating that 50% of neurons are randomly dropped during training.

Table 1: Key parameter settings for LSTM model training.

To mitigate over-fitting, techniques like dropout and regularization are employed. Dropout randomly discards the output of certain neurons during training, enhancing the model's generalization capability. Regularization prevents over-fitting by constraining network parameters. The algorithm dynamically adjusts each parameter's learning rate using the first and second-moment estimations of the gradient. The specific training formula is as follows:

$$\hat{m}_t = \frac{m_t}{1-\mu^t} \tag{10}$$

$$\hat{n}_t = \frac{n_t}{1-\upsilon^t} \tag{11}$$

$$\Delta\theta_t = -\frac{\hat{m}_t}{\sqrt{\hat{n}_t}+\varepsilon} \times \eta \tag{12}$$

Among them, m_t n_t are the first-order estimation and second-order estimation of gradient, respectively; $\hat{m}_t$ and $\hat{n}_t$ are corrections to them, which are approximately unbiased estimates. Through repeated iterative training and using a verification set to monitor the performance of the model, an LSTM model that can accurately extract and classify music features is finally obtained.

4 EXPERIMENTAL DESIGN AND IMPLEMENTATION

The environment configuration of this experiment includes hardware and software, as shown in Table 2.

Category	*Configuration Item*	*Detailed Description*
Hardware	Processor	Intel Xeon E5-2680 v4
	Memory	64GB
	GPU	NVIDIA Tesla P100
Software	Operating System	Ubuntu 18.04
	Programming Language	Python 3.7
	Deep Learning Framework	TensorFlow 2.3
	Data Processing Libraries	NumPy, Pandas

Table 2: Experimental environment configuration table.

The dataset used in this experiment is a music melody dataset that includes multiple music styles and genres. These melody data exist in the form of audio files (such as MP3, and WAV formats), with each file representing a unique piece of music. The dataset covers a wide range of music genres, including but not limited to classical, jazz, rock, pop, electronic, etc., to ensure that the model can learn the characteristics of different styles of music. The dataset contains thousands of audio files, each representing a musical melody with a duration ranging from a few seconds to a few minutes. Each audio file comes with metadata, including but not limited to file name, artist, genre, release year, etc. These metadata are crucial for subsequent label assignment and classification tasks. The dataset assigns at least one label to each audio file to represent its genre, emotion, or other musical attributes. These labels are the target variables for subsequent classification tasks. For the data set division experiment, the dataset is partitioned into three segments: training set, verification set, and test set, with respective proportions of 70%, 15%, and 15%. This division strategy ensures ample data for the model to learn during training, allows for model parameter adjustment using the verification set, and ultimately evaluates the model's generalization ability with the test set. Before division, datasets undergo preprocessing steps, which include outlier removal, standardization, and more, to guarantee data quality and consistency.

The experimental flow includes the following steps: data preprocessing and division; Constructing the LSTM model and setting initial parameters; Model training is carried out, during which parameters are adjusted by using a verification set; Evaluate model performance using a test set; The experimental results are analyzed, and the models are optimized and compared.

5 EXPERIMENTAL RESULTS AND ANALYSIS

For this experiment, Accuracy, Recall, Precision, and F1 scores serve as the primary evaluation metrics, providing a comprehensive assessment of the model's performance in classification tasks. Figure 3 displays the LSTM model's scores for each metric.

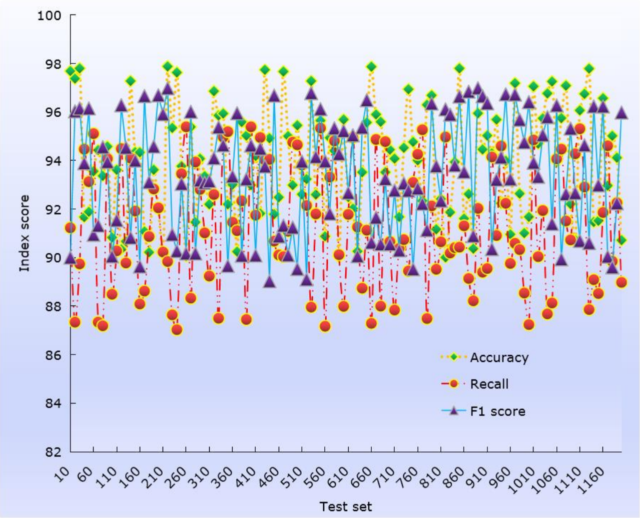

Figure 3: Index score of LSTM model.

Figure 3 visually illustrates the "exponential scores" of the LSTM model in classification tasks for three key evaluation metrics: accuracy, F1 score, and recall rate. The horizontal axis of Figure 3 represents the test set, and the vertical axis represents the exponential score. From the graph, it can be seen that the LSTM model achieved a high score in accuracy. This indicates that the model can accurately classify the music melody samples in the test set into the correct genre or emotion category. Meanwhile, the score of the F1 value is relatively high, which is the harmonic mean of accuracy and recall. A high F1 value means that the model has achieved a good balance between accuracy and recall, which can accurately identify positive samples and recall most of them. To further validate the LSTM model's effectiveness, the experiment compares it with CNN and RNN models, demonstrating its superiority in all indices, notably in handling sequence data. The findings are presented in Figures 4 and 5.

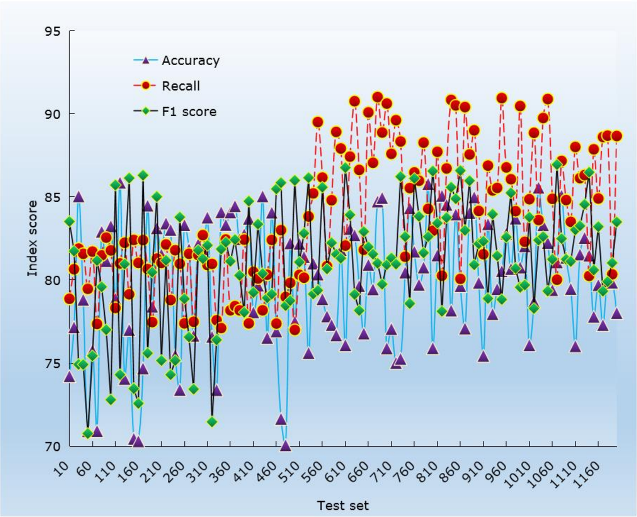

Figure 4: Index score of CNN model.

Figure 4 clearly shows the "exponential score" of the LSTM model in classification tasks for three key evaluation metrics: accuracy, F1 score, and recall rate. The horizontal axis of Figure 4 represents the test set, and the vertical axis represents the exponential score. Figure 4 provides valuable insights into the performance evaluation of the LSTM model in classification tasks. By analyzing the scores of each evaluation metric in-depth, we can better understand the advantages and potential room for improvement of the model.

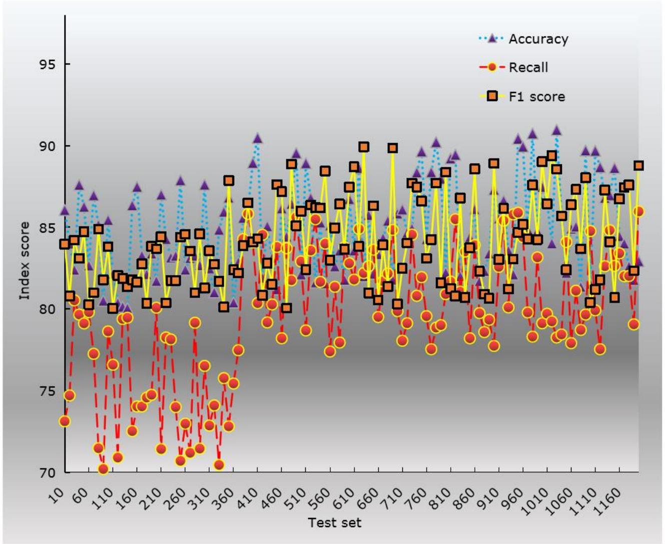

Figure 5: Index score of RNN model.

Figure 5 shows the scores of the LSTM model for accuracy, F1 score, and recall in classification tasks, with the test set as the horizontal axis and the index score as the vertical axis. Although the recall score is good, it is slightly lower compared to accuracy and F1 score. This means that in some cases, the model fails to recognize all samples belonging to the target category. That is, some positive samples are incorrectly classified as other categories or not recognized by the model. The results show that the CNN model achieves an Accuracy of 80.9%, Recall of 86.4%, and F1 score of 83.5% on the test set, while the RNN model attains an Accuracy of 85.6%, Recall of 82.3%, and F1 score of 85.0%. However, the LSTM surpasses them with an Accuracy of 95.7%, Recall of 92.8%, and F1 score of 94.9%, indicating its superior classification performance. Figure 6 shows the classification Precision of each model.

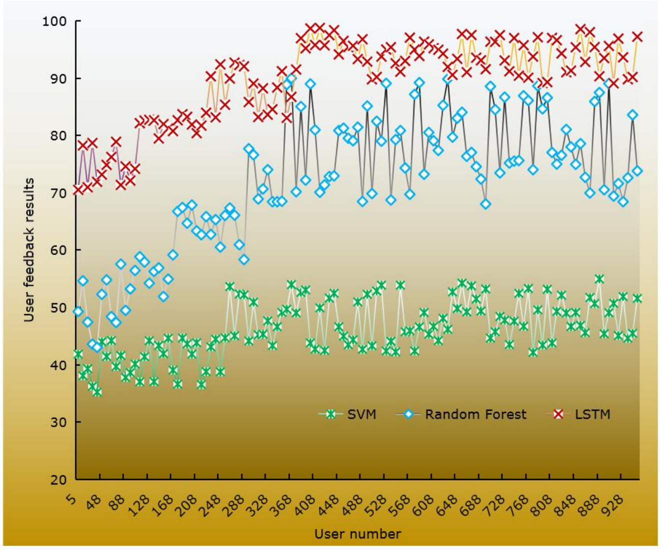

Figure 6: Classification Precision of each model.

The horizontal axis of Figure 6 represents the number of users, and the vertical axis represents the feedback results from users. Figure 6 shows the classification accuracy results of support vector machine, random forest, and LSTM. Among them, LSTM has the highest classification accuracy, followed by random forest, and support vector machine has the lowest classification accuracy. From the graph, it can be seen that the LSTM model performs the best in classification accuracy, thanks to its ability to capture long-term dependencies in sequential data, especially for data with time-series characteristics such as music melodies. The high classification accuracy of LSTM proves its advantages in handling complex classification tasks, especially in scenarios where understanding sequence contextual information is required.

Compared to SVM and RF models, the LSTM model demonstrates superior classification Precision, achieving approximately 0.96. Analysis reveals that while the model exhibits high Accuracy in certain data categories, it performs slightly poorly in others, potentially due to unbalanced data distribution. Furthermore, the balance between Recall and Precision indicates a need to adjust the threshold based on practical application requirements. The experiment additionally examines the impact of various features on classification performance, as illustrated in Figure 7.

The horizontal axis in Figure 7 represents the contribution value, and the vertical axis represents the test samples. We tested text features, time series features, and image features separately. The results in Figure 7 indicate that time series features and text features have the most significant impact on improving model performance, while image features have a certain contribution, but are relatively small. This indicates that in future research, we can focus more on the effective utilization of time series and text features.

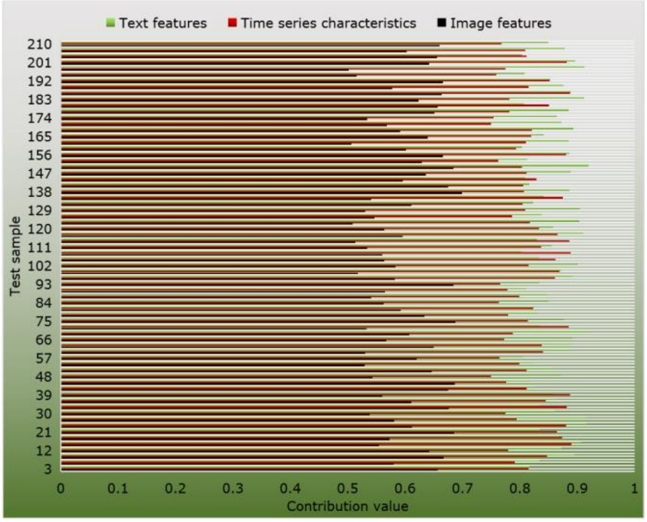

Figure 7: Influence of different features on classification performance.

The experiment also discusses the influence of multi-modal feature fusion on the model performance, such as Table 3, Table 4 and Table 5.

Feature Type	*Model Accuracy*
Single-Modal Features	79.2%
Multi-Modal Feature Fusion	95.5%

Table 3: Impact of single-modal and multi-modal feature fusion on model accuracy.

In this table, the accuracy of the model is 79.2% when the single modal feature is used, but it is improved to 95.5% after multi-modal feature fusion.

Feature Type	*Model Recall*
Single-Modal Features	70.1%
Multi-Modal Feature Fusion	92.5%

Table 4: Impact of single-modal and multi-modal feature fusion on model recall.

In this table, the Recall of the model is 70.1% when the single modal feature is used, but it is improved to 92.5% after multi-modal feature fusion.

Feature Type	*Model F1 Score*
Single-Modal Features	0.81
Multi-Modal Feature Fusion	0.94

Table 5: Impact of single-modal and multi-modal feature fusion on model f1 score.

This table indicates that the model's F1 score is 0.81 when using a single modal feature, but it increases to 0.94 after multi-modal feature fusion. A comparison reveals that multi-modal feature fusion significantly enhances the model's Accuracy, Recall, and F1 scores, emphasizing the importance of effectively integrating different modes' characteristics in scenarios with abundant multimodal data to improve model performance.

6 CONCLUSIONS

Music feature recognition is one of the most fundamental and active research topics in the field of music information retrieval. Music feature recognition plays a fundamental role in the context of music retrieval and can be used for music engines on websites and devices to manage and tag music content. The classification of music genres is different from the problem of image classification, as music genres contain many highly diverse intermediate features with different levels of abstraction. They are usually distributed in local and repetitive audio segments. Music genres are comprehensive descriptions of local content at different levels. Deep learning-based methods focus on global feature learning, making decisions on features at the same level as typical image classification. This ignores the different hierarchical local characteristics of audio and their connections. Deep learning has achieved great success in computer vision and natural language processing, and representation learning methods based on deep learning are gradually being applied to music feature recognition. However, there are still some limitations to the methods currently used for MGR. Therefore, this article proposes a feature fusion algorithm based on multimodal data classification feature encoding for music feature recognition tasks. And it also provides new ideas for addressing the limitations of current music feature recognition methods. To fully explore the local features of different abstract levels in music genres and learn their long-term dependencies, this paper is inspired by the LSTM model mechanism to design a feature encoding network for music feature recognition. In the final model integration, this paper uses the LSTM model to learn the interaction relationship between different features in the early stage for feature-level data fusion, rather than combining decisions from independent classifiers. The experimental results show that the feature fusion algorithm based on multi-level local feature encoding has achieved leading performance on the dataset, demonstrating its good application scenarios.

Feiyan Yang, https://orcid.org/0009-0001-1424-1185

REFERENCES

[1] Bishop, L.; Cancino, C.-C.; Goebl, W.: Moving to communicate, moving to interact: Patterns of body motion in musical duo performance, Music Perception: An Interdisciplinary Journal, 37(1), 2019, 1-25. https://doi.org/10.1525/mp.2019.37.1.1

[2] Calilhanna, A.: Ogene Bunch music analyzed through the visualization and sonification of beat-class theory with ski-hill and cyclic graphs, The Journal of the Acoustical Society of America, 148(4), 2020, 2697. https://doi.org/10.1121/1.5147469

[3] Chaturvedi, V.; Kaur, A.-B.; Varshney, V.; Garg, A.; Chhabra, G.-S.; Kumar, M.: Music mood and human emotion recognition based on physiological signals: a systematic review, Multimedia Systems, 28(1), 2022, 21-44. https://doi.org/10.1007/s00530-021-00786-6

[4] Dong, Y.; Yang, X.; Zhao, X.: Bidirectional convolutional recurrent sparse network (bcrsn): an efficient model for music emotion recognition, IEEE Transactions on Multimedia, 21(12), 2019, 3150-3163. https://doi.org/10.1109/TMM.2019.2918739

[5] Dorris, J.-L.; Chang, K.; Mclaughlin, D.-J.: Project unmute: a digital music program delivered by adolescent musicians to older adults with cognitive decline, Journal of Intergenerational Relationships, 20(4), 2022, 493-501. https://doi.org/10.1080/15350770.2022.2086958

[6] Dotov, D.; Bosnyak, D.; Trainor, L.-J.: Collective music listening: movement energy is enhanced by groove and visual social cues, Quarterly Journal of Experimental Psychology, 74(6), 2021, 1037-1053. https://doi.org/10.1177/1747021821991793

[7] Ge, M.; Tian, Y.; Ge, Y.: Optimization of computer-aided design system for music automatic classification based on feature analysis, Computer-Aided Design and Applications, 19(S3), 2021, 153-163. https://doi.org/10.14733/cadaps.2022.S3.153-163

[8] Georges, P.; Seckin, A.: Music information visualization and classical composers discovery: an application of network graphs, multidimensional scaling, and support vector machines, Scientometrics, 127(5), 2022, 2277-2311. https://doi.org/10.1007/s11192-022-04331-8

[9] Gorbunova, I.-B.; Plotnikov, K.-Y.: Music computer technologies in education as a tool for implementing the polymodality of musical perception, Musical Art and Education, 8(1), 2020, 25-40. https://doi.org/10.31862/2309-1428-2020-8-1-25-40

[10] Hizlisoy, S.; Yildirim, S.; Tufekci, Z.: Music emotion recognition using convolutional long short term memory deep neural networks, Engineering Science and Technology an International Journal, 24(3), 2020, 760-767. https://doi.org/10.1016/j.jestch.2020.10.009

[11] Huang, Z.; Jia, X.; Guo, Y.: State-of-the-art model for music object recognition with deep learning, Applied Sciences, 9(13), 2019, 2645. https://doi.org/10.3390/app9132645

[12] Jiang, J.: Signal feature extraction of music melody based on deep learning, Traitement du Signal, 39(6), 2022, 2203-2209. https://doi.org/10.18280/ts.390635

[13] Kumaraswamy, B.: Optimized deep learning for genre classification via improved moth flame algorithm, Multimedia Tools and Applications, 81(12), 2022, 17071-17093. https://doi.org/10.1007/s11042-022-12254-y

[14] Li, S.; Luo, Q.; Qiu, L.: Optimal pricing model of digital music: subscription, ownership or mixed? Production and Operations Management, 29(3), 2020, 688-704. https://doi.org/10.1111/poms.13131

[15] Liao, Y.; Gui, Z.: An intelligent sparse feature extraction approach for music data component recognition and analysis of hybrid instruments, Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology, 45(5), 2023, 7785-7796. https://doi.org/10.3233/JIFS-231290

[16] Maba, A.: Computer-aided music education and musical creativity, Journal of Human Sciences, 17(3), 2020, 822-830. https://doi.org/10.14687/jhs.v17i3.5908

[17] Panda, R.; Malheiro, R.; Paiva, R.-P.: Audio features for music emotion recognition: a survey, IEEE Transactions on Affective Computing, 14(1), 2020, 68-88. https://doi.org/10.1109/TAFFC.2020.3032373

[18] Pei, Z.; Wang, Y.: Analysis of computer aided teaching management system for music appreciation course based on network resources, Computer-Aided Design and Applications, 19(S1), 2021, 1-11. https://doi.org/10.14733/cadaps.2022.S1.1-11

[19] Xiang, H.: The collection of theater music data and genre recognition under the Internet of things and deep belief network, The Journal of Supercomputing, 78(7), 2022, 9307-9325. https://doi.org/10.1007/s11227-021-04261-x