International Journal of Computational Intelligence Systems (2024) 17:99
https://doi.org/10.1007/s44196-024-00489-6

RESEARCH ARTICLE

MMD-MII Model: A Multilayered Analysis and Multimodal Integration Interaction Approach Revolutionizing Music Emotion Classification

Jingyi Wang[1] · Alireza Sharifi[2] · Thippa Reddy Gadekallu[3,4,5,6,7] · Achyut Shankar[8,9,10]

Received: 30 December 2023 / Accepted: 26 March 2024 / Published online: 22 April 2024

Abstract
Music plays a vital role in human culture and society, serving as a universal form of expression. However, accurately classifying music emotions remains challenging due to the intricate nature of emotional expressions in music and the integration of diverse data sources. To address these challenges, we propose the Multilayered Music Decomposition and Multimodal Integration Interaction (MMD-MII) model. This model employs cross-processing to facilitate interaction between audio and lyrics, ensuring coherence in emotional representation. Additionally, we introduce a hierarchical framework based on the music theory, focusing on the main and chorus sections, with the chorus processed separately to extract precise emotional representations. Experimental results on the DEAM and FMA datasets demonstrate the effectiveness of the MMD-MII model, achieving accuracies of 49.68% and 49.54% respectively. Compared with the existing methods, our model outperforms in accuracy and $F1$ scores, offering promising implications for music recommendation systems, healthcare, psychology, and advertising, where accurate emotional analysis is essential.

Keywords Multimodal · Music emotion classification · Emotion analysis · Music structure analysis · Deep learning · Music feature extraction

1 Introduction

Music holds a profound significance in human culture and life, exerting a profound influence on a global scale. Spanning from classical melodies to contemporary pop beats, and from traditional folk tunes to cutting-edge electronic compositions, music embodies a vast spectrum of styles, serving as a powerful medium for inspiration and emotional expression [1]. Beyond its artistic essence, music serves as a multifaceted tool for entertainment, social cohesion, cultural preservation, and psychological well-being. In recent decades, propelled by the rapid advancements in multimedia technology, music has become increasingly accessible, integrating seamlessly into people's daily lives, serving as

✉ Jingyi Wang
tianyue1108@163.com

Alireza Sharifi
a_sharifi@sru.ac.ir

Thippa Reddy Gadekallu
thippareddy@ieee.org

Achyut Shankar
ashankar2711@gmail.com

[1] School of Music, Jiangxi Normal University, Nanchang 330027, Jiangxi, China

[2] Department of Surveying Engineering, Faculty of Civil Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran

[3] Zhongda Group, Haiyan County, Jiaxing 314312, Zhejiang, China

[4] Department of Electrical and Computer Engineering, Lebanese American University, Byblos, Lebanon

[5] School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, India

[6] College of Information Science and Engineering, Jiaxing University, Jiaxing 314001, China

[7] Division of Research and Development, Lovely Professional University, Phagwara, India

[8] WMG, University of Warwick, Coventry, UK

[9] Center of Research Impact and Outcome, Chitkara University, Punjab 140401, India

[10] School of Computer Science Engineering, Lovely Professional University, Punjab 144411, Phagwara, India

a source of leisure and a means of emotional regulation [2]. Central to the essence of music is its capacity to evoke and convey emotions, serving as a conduit for profound emotional experiences. Whether evoking feelings of joy, melancholy, anticipation, or tranquility, music possesses the remarkable ability to elicit emotional resonance within listeners, offering a medium for emotional exploration and catharsis. Consequently, the study of emotions in music has emerged as a pivotal area of research within the realms of music psychology and computer science.

However, despite significant advancements, emotion classification tasks in music analysis still confront several challenges. Traditional methods often rely solely on audio data for classification, overlooking the rich potential offered by other multimodal sources such as lyrics, music videos, and social media comments [3]. Moreover, existing classification models may struggle to effectively capture the complexity of emotional expressions and nuances inherent in music, given the multidimensional nature of emotions, which extend beyond simple binary distinctions like pleasure or sadness.

In response to these challenges, multimodal approaches offer a promising avenue for enhancing emotion classification tasks in music analysis. By harnessing insights from diverse multimodal data sources, including audio, text, images, and videos, researchers can cultivate a more holistic understanding of the emotional landscapes embedded within music [4]. Multimodal music emotion classification not only enhances classification accuracy but also enables a nuanced portrayal of the intricate emotional tapestry woven within musical compositions. This advancement holds transformative potential, empowering music recommendation systems to tailor experiences to the emotional needs of listeners, while also fostering applications in domains such as healthcare, psychology, and advertising.

In traditional research on music emotion classification, researchers extensively utilize various audio feature extraction techniques to obtain information from audio signals for use in emotion classification. Common traditional audio feature extraction techniques include MFCC and Spectral Centroid. MFCC, in particular, is a classical method for audio feature extraction widely used in music and speech processing [5]. It simulates the way human ears perceive sound to extract spectral information from audio signals. Spectral Centroid represents the central position of the spectrum and is used to describe the pitch attributes of the audio [6]. This feature is often employed in music emotion classification to aid in distinguishing pitch variations under different emotional states. In addition to traditional audio feature extraction techniques, some advanced deep-learning models have garnered significant attention. For example, CNNs, are not only used in image processing but also extensively applied in audio processing [7]. They can effectively capture local features in audio signals, making CNNs highly valuable in music emotion classification. In addition, RNNs, which are recurrent neural networks specialized in handling time-series data, are helpful in capturing time-related features in audio signals, particularly when describing emotional changes in music.

Furthermore, the use of pretrained deep learning models, such as BERT or GPT, has led to significant improvements in music emotion classification [8]. This is because these models have undergone extensive training in processing text data, possessing strong semantic understanding and representation learning capabilities that contribute to a better understanding of emotional information in music [9]. However, these approaches do not take into account the often-existing consistency of emotions between lyrics and melody. In addition, music possesses natural structural information, and these approaches have not considered the inherent structure of music (such as verse-chorus), which is highly effective and necessary for music emotion analysis [10].

Drawing from these insights, we introduce the MMD-MII (Multilayered Music Decomposition and Multimodal Integration Interaction) model, a cutting-edge multimodal framework designed to enhance music emotion recognition and analysis. Our model incorporates the inherent structural elements of music, specifically focusing on the verse and chorus sections, while facilitating interaction between modalities during processing. Upon input, music undergoes cross-processing, enabling seamless interaction between audio and lyrics to maintain emotional coherence. In addition, we establish a hierarchical framework based on the theory of music's verse and chorus, conducting separate analysis on the chorus section to extract precise emotional representations. The overarching objective of the MMD-MII model is to integrate audio, lyrics, and other multimodal data sources at multiple levels, while considering the intrinsic structure of music to significantly elevate music emotion recognition and analysis performance. Through interactive processing and hierarchical analysis, our model offers unparalleled accuracy in capturing emotional information in music, effectively catering to the diverse emotional needs of audiences. Furthermore, the MMD-MII model holds immense potential for transformative applications in fields such as music recommendation, advertising, and psychology, promising to reshape the landscape of multimodal emotion analysis in music research.

The article primarily makes three contributions:

- This article first introduces a hierarchical music analysis framework for analyzing the structure of music. Then, it constructs a novel multimodal interaction framework, extracting the current emotion vector at each time step, and further fusing and updating these emotion vectors to ensure emotional consistency among various modalities.

- The MMD-MII model not only integrates multimodal data but also places a specific focus on the intrinsic structure of music, including aspects like the verse and chorus. Through the hierarchical framework, we can more accurately extract and analyze emotions in different parts of the music, facilitating a deeper understanding of emotional expression in music.
- The MMD-MII model introduces emotion vectors and designs emotion LSTM cell units to effectively capture emotional information in music, especially when dealing with datasets featuring four different emotion labels. This provides a more accurate and in-depth approach to emotion analysis.

In the rest of this paper, we present recent related work in Sect. 2. Section 3 introduces our proposed methods. Section 4 showcases the experimental part. Section 5 contains the conclusion.

2 Related Work

2.1 Research on Lyric Text Processing

In the research of music emotion classification, lyric text processing is a crucial step that helps models better understand emotional information within music. Within this research domain, various deep learning models have emerged for processing lyric text, aimed at enhancing the performance of music emotion classification. These models include BERT, GPT-3.5, XLNet, RoBERTa, and DistilBERT.

BERT (Bidirectional Encoder Representations from Transformers) is a bidirectional pretrained model that, through deep learning, can delve deeper into the understanding of emotional content within the lyric text [11]. GPT-3.5, with its enormous parameter count, excels in text generation and comprehension, offering robust support for lyric emotion analysis and lyric generation [12]. XLNet employs a different pretraining approach, aiding in capturing a more comprehensive view of dependencies in text and providing additional angles for emotional understanding [13]. RoBERTa represents an enhancement of BERT, achieved through a larger dataset and extended training duration, resulting in improved performance and more precise emotional representations [14]. Furthermore, DistilBERT is a lightweight version of BERT, offering computational efficiency while still performing well in lyric text processing [14].

These deep learning models provide a diverse set of tools for lyric text processing, enabling the automatic extraction of emotional features from lyrics without manual intervention. Researchers can choose an appropriate model based on the task requirements and computational resources to enhance the accuracy and performance of music emotion classification. These models play a pivotal role in the emotional analysis of lyric text.

3 Research Based on Audio Feature Extraction

In the research of music emotion classification, melodic audio processing plays a crucial role in extracting melodic information from audio signals to enhance emotion classification performance. Several novel audio feature extraction methods have emerged in this research domain [15]. For instance, Deep Chroma is a deep learning-based audio feature extraction method that focuses on capturing harmony, chords, and melodic information in audio. This model employs a combination of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to analyze the temporal features of audio signals and extract melodic information from music [16]. Deep learning methods like Deep Chroma have the advantage of automatic feature learning, aiding in a more accurate capture of emotional elements within audio. In addition, WaveNet, originally designed for audio waveform generation, has proven to be valuable in audio analysis. WaveNet can model audio signals at high resolutions, capturing fine-grained audio features and melodic variations, providing more informative features for music emotion classification [17]. Furthermore, transfer learning methods have gained prominence in audio processing. They involve the use of pretrained audio models such as VGGish or OpenL3, which offer significant performance improvements in music emotion classification. These models have undergone extensive pretraining on large audio datasets and can automatically extract audio features, eliminating the need for manual feature engineering [18].

These novel audio feature extraction methods offer diverse tools for melodic audio processing, enabling researchers to better understand and analyze emotional elements within music. They allow for a more accurate capture of emotional information within audio signals, thereby improving the performance of music emotion classification. Researchers can select appropriate models based on task requirements and dataset characteristics, unlocking greater potential in the field of music emotion classification [19].

3.1 The Multimodal Models for Music Emotion Classification

In the field of music emotion classification, multimodal models have made significant progress in handling different modalities of data, including audio, lyrics, images, and more. For example, MuSeNet is designed to fuse audio and lyric information to more accurately capture emotions in music [20]. MuSeNet employs a deep neural network

structure capable of processing both audio and text data. The model consists of two key components: a multimodal encoder and an adaptive module. The multimodal encoder is used to extract feature representations from audio and lyrics, while the adaptive module dynamically learns the weights between different modalities to achieve better performance. MuSeNet's uniqueness lies in its ability to effectively integrate information from different modalities, thus improving the accuracy and performance of music emotion classification. Another model, FusionNet, combines audio and lyric information [21]. It uses deep convolutional neural networks (CNN) and recurrent neural networks (RNN) structures to process different modal data. This model also introduces fusion strategies to gradually combine modality information for music emotion classification. In addition, MuSeCAR combines audio, lyrics, and emotions to gain a more comprehensive understanding of music emotions [22]. This model incorporates deep learning and knowledge graph techniques by combining multimodal data with an emotional knowledge graph, enabling deeper analysis of emotional content. What sets MuSeCAR apart is its ability not only to predict emotions but also to explain the reasons behind those emotions, providing a more in-depth emotional analysis.

These multimodal models provide powerful tools for music emotion classification, with the potential to more accurately capture emotional elements in music by integrating information from different modalities, thereby enhancing the performance of emotion classification. The ongoing development and innovation of these models will further drive research and applications in the field of music emotion classification.

4 Methodology

We propose the MMD-MII multimodal music emotion classification model. Firstly, we utilize VGGish to extract audio features and ALBERT to extract lyric features. We then introduce the inherent structure of music (verse and chorus) into the overall model framework, enabling interactions between modalities during processing. The goal is to enhance music emotion recognition and analysis. The model diagram is shown in Fig. 1. Once music is input, it passes through a module called the "cross-processing" module. Within this module, audio and lyrics interact to ensure emotional consistency. Simultaneously, we employ a hierarchical framework based on the theory of music's verse and chorus. When the music reaches the chorus section, we process it separately, extracting more accurate emotional representations.

4.1 VGGish Module

VGGish is a widely employed deep learning model in the field of audio, specialized for audio feature extraction and audio content analysis [23]. Firstly, VGGish adopts a convolutional neural network (CNN) architecture, bearing some similarities to the VGG models used in the visual domain. The core task of this model is to extract high-level audio features from audio spectrograms, which can be utilized for various audio analysis tasks. Secondly, VGGish takes short time segments of audio signals as input and employs convolution and pooling layers to analyze these audio features. The model's output is a fixed-length vector representing a high-level representation of the audio segment. These embedding vectors can be used

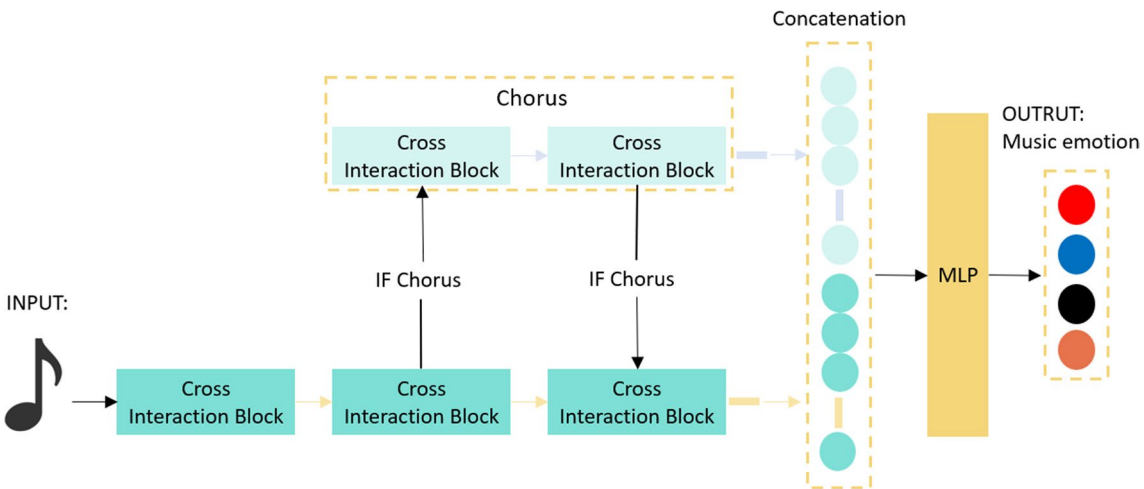

Fig. 1 Overall flow chart of the model

in subsequent tasks such as audio classification, music emotion analysis, and environmental sound recognition.

VGGish plays a crucial role in our model by providing essential support for the processing and feature extraction of audio data. It enables our model to better understand and analyze audio content, thus enhancing both its performance and versatility. Figure 2 depicts the flowchart of the VGGish Module.

4.2 ALBERT Module

In our framework, ALBERT is applied to extract features from lyrics text, facilitating the model's enhanced comprehension and analysis of lyrical content [7]. The subsequent points elucidate ALBERT's pivotal role in extracting lyric features:

Initially, ALBERT undergoes pretraining on extensive textual datasets, enabling it to acquire comprehensive representations of textual information. This pretrained model is adept at extracting general text characteristics, encompassing those pertinent to lyrics text.

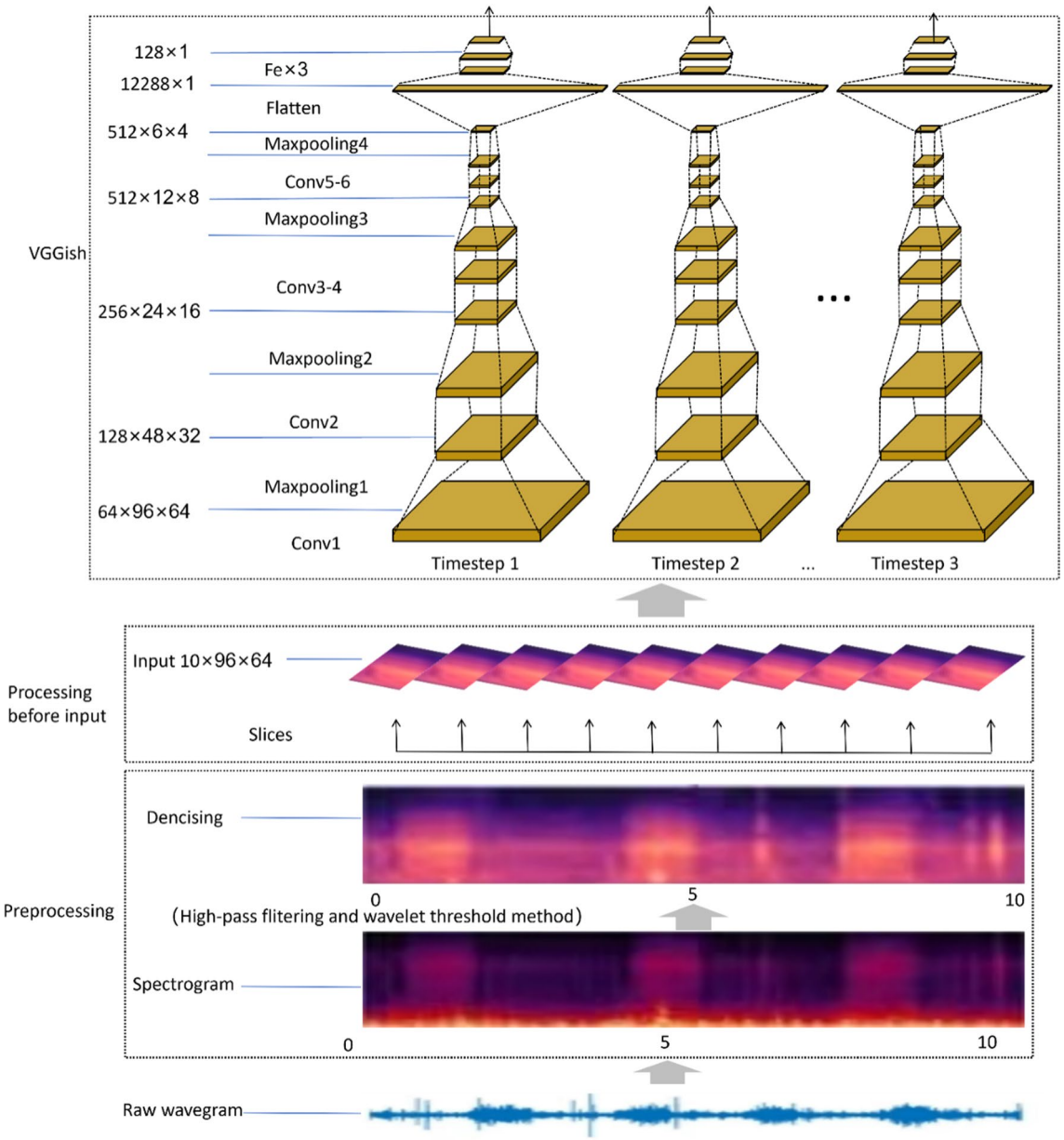

Fig. 2 The VGGish structural unit

Subsequently, upon inputting lyrics text into the ALBERT model, it undergoes conversion into text embedding vectors. These vectors encapsulate the semantic essence and structural composition of the lyrics, constituting high-dimensional feature representations that frequently encapsulate rich semantic information.

Furthermore, ALBERT exhibits context awareness, possessing the capacity to discern intricate relationships between words and contextual cues. This contextual comprehension holds paramount importance for processing lyrics text, given its propensity for harboring nuanced implicit meanings and emotional nuances.

Moreover, in our model, ALBERT's capabilities extend beyond mere semantic analysis. It actively integrates contextual information and emotional nuances from lyrics text, thereby enriching the model's understanding of the lyrical content's emotional depth and complexity.

Conclusively, ALBERT serves as a fundamental component in our framework for extracting features from lyrics text, thereby amplifying the model's proficiency in comprehending and analyzing the emotional dimensions of music lyrics.

ALBERT is a deep learning model used for extracting features from lyrics text. Through pretraining, it can convert lyrics text into meaningful high-dimensional feature representations, which contribute to a better understanding and analysis of the emotional content in music lyrics. This provides strong support for music emotion classification tasks.

Figure 3 displays the network structure of the ALBERT model.

4.3 Cross Processing Module

The Cross Processing Module is a module designed for handling multimodal information. It is configured to facilitate interaction between lyrics and melody, and its output is an emotion feature vector. This emotion feature vector is jointly constrained by the melody and lyrics modules. Furthermore, this emotion feature vector is subsequently input into the next set of lyric–melody pairs to ensure emotional consistency between melody and lyrics. The specific structure is depicted in Fig. 4. This module is primarily utilized for processing lyric–melody pairs, enabling interaction between the two modalities.

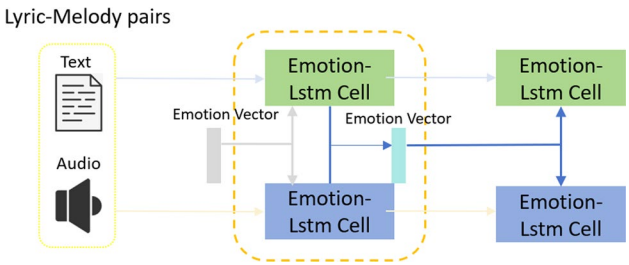

Fig. 4 The Cross Processing Module network structure

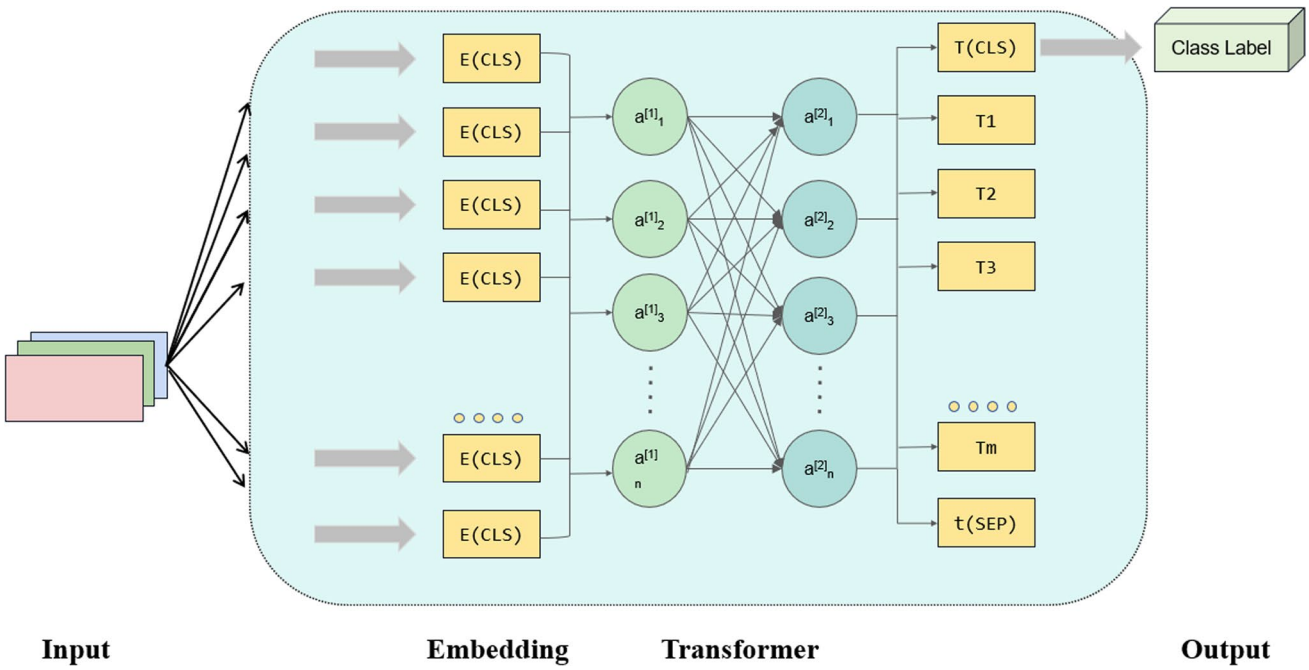

Fig. 3 The basic ALBERT network structure

In Fig. 4, we can see that at any given moment, the input to the Emotion–LSTM includes not only the lyric–melody pair but also a vector referred to as the "emotion vector." This vector is determined after the previous interaction between the lyric and melody (with the first emotion vector originating from random initialization). When the current lyric–melody pair enters the paired Emotion-LSTM, they each produce an emotion vector. These emotion vectors generated from both modalities then interact to create a new emotion vector, fusing information from both channels. This approach ensures that the two modalities under this vector maintain a consistent emotional state with respect to the previous moment's lyric–melody pair and prevents emotions between the two channels from being independent.

4.4 Emotion–LSTM model

The Emotion–LSTM model represents an enhancement of the traditional LSTM architecture, with a particular focus on the incorporation of emotion vectors as inputs. This study aims to investigate the precise allocation of emotional weighting within the context of lyrical content and melody [7]. To accomplish this, a novel two-polarity emotion vector is introduced, designed to partition emotions into distinct categories. The upper segment of the vector corresponds to positive emotions, signifying heightened states of joy and euphoria. The middle portion is dedicated to neutral emotions, while the lower section encapsulates increasingly melancholic and negative emotions. The Emotion–LSTM cell unit, delineated in Fig. 5, serves as the cornerstone of this framework, providing a detailed overview of its inner workings. The incorporation of the emotion vector empowers the model to adeptly capture and comprehend the intricate emotional nuances conveyed through music. Notably, the dataset used in this research comprises four distinct emotional classes. It encompasses not only the polar emotional states of sadness and happiness but also embraces two additional nuanced emotional categories: tranquility and healing.

As depicted in Fig. 5, the Cum–Sum process involves the integration of both the historical emotion vector and the current input emotion vector, akin to a single interaction between past and present emotions. The historical emotion vector primarily aims to retain intense emotional information, while the current emotion vector focuses on refreshing relatively weaker emotions. The resultant emotion vector undergoes continuous iteration alongside the historical emotion vector. This iterative process facilitates the determination of which emotional levels within the song should be retained and which previously utilized emotions necessitate updating. The hierarchical emotion vector maintains its connection with the cell state "C" and the hidden state "h" within the Long Short-Term Memory (LSTM) framework, exerting a guiding influence on updates to the cell state "C" and the hidden state "h." The specific operations will be elucidated and elaborated upon through subsequent mathematical formulations.

Input gate:

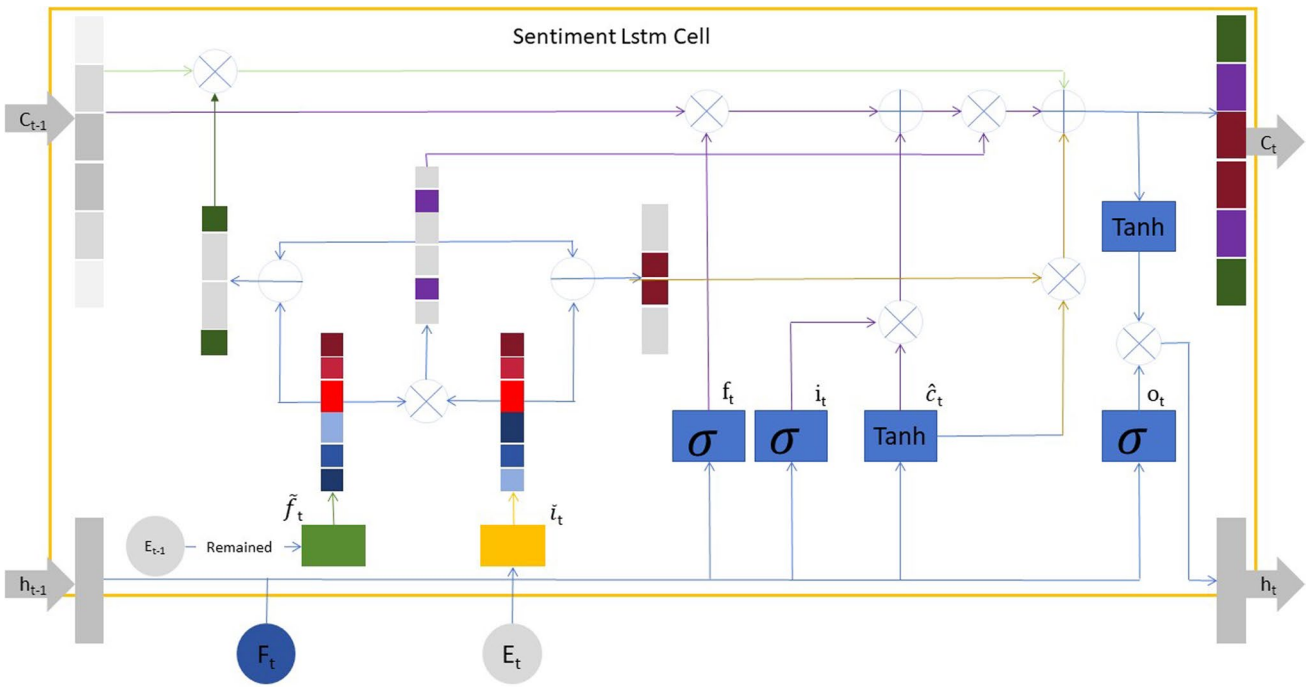

Fig. 5 The Emotion–LSTM network structure

$$i_t = \sigma(W_i F_t + U_i h_{t-1} + b_i)$$

Forget gate:

$$f_t = \sigma(W_f F_t + U_f h_{t-1} + b_f)$$

Output gate

$$o_t = \sigma(W_o F_t + U_o h_{t-1} + b_o)$$

Candidate cell state:

$$\widetilde{C}_t = \tanh(W_c F_t + U_i h_{t-1} + b_c)$$

Emotion forget gate (negative emotion part):

$$\widetilde{f}_t^{[0:l/2]} = \overline{\text{cumsum}}\left(\text{softmax}\left(E_{t-1}\left[0;\frac{l}{2}\right]\right)\right)$$

Emotion forget gate (positive emotion part):

$$\widetilde{f}_t^{[l/2:l]} = \overline{\text{cumsum}}\left(\text{softmax}\left(E_{t-1}\left[\frac{l}{2}:l\right]\right)\right)$$

Emotion forget gate:

$$\widetilde{f}_t = \text{concat}\left[\widetilde{f}_t^{\left[0:\frac{l}{2}\right]}, \widetilde{f}_t^{\left[\frac{l}{2}:l\right]}\right]$$

Emotion input gate (negative emotion part):

$$\widetilde{i}_t^{[0:l/2]} = \overline{\text{cumsum}}\left(\text{softmax}\left(E_{t-1}\left[0:\frac{l}{2}\right]\right)\right)$$

Emotion input gate (positive emotion part):

$$\widetilde{i}_t^{[l/2:l]} = \overline{\text{cumsum}}\left(\text{softmax}\left(E_{t-1}\left[\frac{l}{2}:l\right]\right)\right)$$

Emotion input gate:

$$\widetilde{i}_t = \text{concat}\left[\widetilde{i}_t^{\left[0:\frac{l}{2}\right]}, \widetilde{i}_t^{\left[\frac{l}{2}:l\right]}\right]$$

Emotion interaction state:

$$w_t = \widetilde{f}_t \cdot \widetilde{i}_t$$

Update memory unit:

$$C_t = w_t \cdot (f_t \cdot c_{t-1} + i_t \cdot \widetilde{C}_t) + (\widetilde{f}_t - w_t) \cdot c_{t-1} + (\widetilde{i}_t - \mathrm{w}_t) \cdot \widetilde{C}_t$$

In this context, 't' represents the current time step, '$t-1$' represents the previous historical time step, 'f_t' represents the elements input at the current time step 't' (which can be audio or text), and 'E_t' represents the emotion vector for the current time step. The emotion vector for audio is determined and updated jointly by 'h' and 'c' within the emotion LSTM, to model emotional intensity. The emotion vector for lyrics can be mapped using a neural network or an emotional dictionary and is determined after embedding. All 'b' terms are bias terms, and 'σ' represents the sigmoid activation function. Similar to a standard LSTM, there are input gates, output gates, and forget gates, as denoted by the formulas, which represent the three main gates in the LSTM. The candidate cell state 'C_t' contains hierarchical information, and unlike a standard LSTM, it undergoes hierarchical updates based on the emotion vector, with updates to the cell state 'C_t' following new rules.

5 Experiment

5.1 Datasets

The experimental section of this study involves two significant music datasets: the DEAM Dataset (Dynamic Emotional Analysis of Music) and the FMA Dataset (Free Music Archive). These two datasets have widespread applications in the fields of music emotion analysis and music research, providing valuable resources and materials for our research.

The DEAM Dataset is a multimodal music dataset designed to assist researchers in delving deeper into the relationship between music and emotions [24]. It comprises a substantial amount of music audio, emotional annotations, and associated metadata. The emotional annotation segment of this dataset meticulously records the emotional states within the audio, allowing researchers to analyze and understand the subtle variations in emotional expression within music compositions. The multimodal nature of the DEAM Dataset, encompassing both audio and emotional annotations, offers profound insights into the realm of music emotion analysis.

The FMA Dataset is a widely used music dataset, featuring a vast collection of audio tracks and related metadata [25]. It offers music of various genres and styles, spanning a wide range from classical to popular music. The accessibility and open nature of the FMA Dataset make it an ideal choice for music classification, music recommendation, and music research. Researchers can access audio materials from the FMA Dataset for experimentation and analysis to support their music research projects.

5.2 Experimental Environment

My experimental setup consists of: Processor, Intel i7-13650 CPU; Graphics Card, NVIDIA GTX 4090; Memory, 32 GB. The software environment is as follows: General-purpose computing architecture CUDA 11.6; GPU acceleration library, CUDNN 9.0; Deep learning framework Pytorch.

The tool we employed for vocal separation is the open-source program Spleeter. Strictly speaking, Spleeter is

also a form of pretrained model written in TensorFlow and achieves its functionality through the utilization of U-Net architecture. In addition, we use the open-source computer program FFmpeg for sentence alignment between lyrics and melody. FFmpeg is an open-source computer program developed for Linux, used for recording, converting digital audio and video, and transforming them into streams. FFmpeg also serves as an audio or video encoder.

5.3 Evaluation Metrics

This experiment used accuracy rate, precision rate (P), recall rate (R), and $F1$ score to evaluate the model's performance.

Accuracy (accuracy rate): accuracy measures the overall correctness of the model's predictions. It is the ratio of correctly predicted samples to the total number of samples in the dataset.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of samples}}$$

Recall (sensitivity or true positive rate): Recall calculates the proportion of positive samples that are correctly identified by the model. It measures the ability of the model to find all the positive samples.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Precision (precision rate): Precision calculates the proportion of positive predictions made by the model that is correct. It measures the model's ability to avoid false positives.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

F-Score ($F1$-Score): F-Score is the harmonic mean of precision and recall. It is useful when both precision and recall are important, and you want to balance their contribution.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

5.4 Experimental Details

In the DEAM dataset, we selected a subset of songs with relatively high play counts, assuming that these songs exhibit a higher degree of emotional consistency with the provided emotion labels. We chose four emotion labels: happiness, sadness, healing, and calm, and collected approximately 6000 songs. After further filtering based on factors such as song length, audio quality, and language, we retained around 4280 songs as our candidate dataset. In the case of the FMA dataset, we collected approximately 5320 songs and eventually narrowed it down to 4605 songs for our candidate dataset after applying similar selection criteria. In Table 1, the specific partitioning of the dataset is presented.

The vast majority of songs follow a natural structure known as the verse–chorus structure. The verse–chorus structure typically involves dividing a song into two primary sections: one section known as the "Verse," which serves to establish the song's background, and the other section referred to as the "Chorus," which is responsible

Table 1 The size of the dataset in each emotion category

	DEAM dataset				FMA dataset			
	Happy	Sad	Calm	Healing	Happy	Sad	Calm	Healing
Training set	800	1007	999	1200	950	780	880	1000
Test set	202	189	350	253	230	205	302	258

Table 2 The performance of the music emotion recognition task using our proposed framework and different baseline methods

Methods	Multimodal?	DEAM dataset				FMA dataset			
		Accuracy %	Precision %	Recall %	$F1$ %	Accuracy %	Precision %	Recall %	$F1$ %
Ding et al. [26]	×	46.78	46.24	45.12	46.02	46.54	45.98	46.04	46.26
Kim et al. [27]	×	45.11	46.64	44.31	46.42	56.85	46.21	46.35	46.72
Catharin et al.[28]	√	47.52	47.96	47.62	47.21	47.35	46.89	46.97	46.91
Pandeya et al. [29]	√	47.91	48.54	47.42	47.46	47.46	47.21	47.31	47.36
Zhao et al. [30]	√	48.23	48.75	47.97	47.97	47.85	47.88	47.25	47.62
Chen et al. [31]	√	47.14	47.55	46.52	46.35	47.55	47.21	47.47	46.52
Medina et al. [32]	√	46.24	47.52	46.32	47.52	47.23	45.22	46.52	47.11
Ours	√	49.68	50.06	50.17	49.84	49.54	49.42	49.55	49.89

for emphasizing and expressing the emotions of the song. We have conducted detailed annotations of the verses and choruses in the music dataset.

5.5 Main Results

In Table 2, we can see that the use of multimodal methods (combining multiple data modalities) generally outperforms single-modal methods. The superiority of multimodal methods lies in their ability to integrate information from different data sources, thus enabling a more comprehensive and in-depth understanding of emotional expression in music. These methods exhibit greater adaptability in the field of music emotion analysis, given the complexity and diversity of emotions in music, which encompass aspects such as sound, lyrics, and melody. Notably, our multimodal method, "Ours," excels on both the DEAM and FMA datasets. Taking the DEAM dataset as an example, our method exhibits an improvement of nearly 2 percentage points in accuracy compared to other methods, with significantly higher precision, recall, and $F1$ scores. These substantial performance improvements clearly demonstrate the superiority of multimodal methods, and highlight the leading position of our multimodal method in the field of music emotion analysis.

This table provides compelling evidence in favor of employing multimodal methods for music emotion analysis. Multimodal methods not only deliver superior performance but also offer researchers and practitioners a deeper and more comprehensive insight, aiding in a better understanding of emotional expression in music [33, 34]. This is of significant value for research and applications in the field of music [34].

5.6 Ablation Experiment

In Table 3, we conducted a series of ablation experiments, specifically focusing on the model's performance on the DEAM and FMA datasets. These experiments aimed to emphasize the importance of different modules, including using only the verse module and using only the chorus module, with our multimodal approach serving as the benchmark. Clear distinctions can be observed: on the DEAM dataset, our multimodal method achieved an accuracy of 49.68%, while using only the verse module or chorus module resulted in accuracies of 46.73% and 47.53%, respectively. On the FMA dataset, our multimodal approach achieved an accuracy of 49.54%, whereas using only the verse module or chorus module yielded accuracies of 46.22% and 47.52%, respectively. These numerical comparisons clearly demonstrate that our multimodal method outperforms single-modal models in terms of accuracy, showing a specific

Table 3 Ablation comparison of different metrics and our model for different music levels in the music emotion recognition task

Model	DEAM dataset				FMA dataset			
	Accuracy %	Precision %	Recall %	$F1$ %	Accuracy %	Precision %	Recall %	$F1$ %
Only The Verse Module	46.73	46.55	45.23	46.33	46.22	45.52	45.66	47.25
Only Chorus Module	47.53	47.23	47.52	47.23	47.52	46.23	46.33	46.23
Ours	49.68	50.06	50.17	49.84	49.54	49.42	49.55	49.89

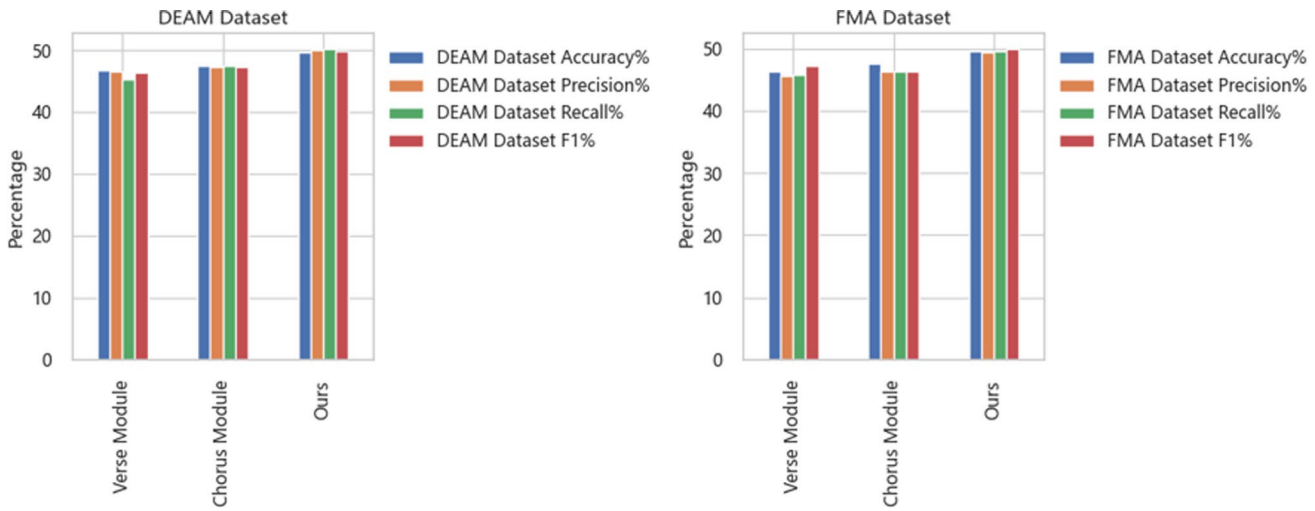

Fig. 6 Comparison of different indicators of different models

improvement of approximately 3–4%. We have visualized the content of the table in Fig. 6.

In Table 4, we present the results of Emotion–LSTM ablation experiments on the DEAM and FMA datasets. These results aim to compare the performance of different recurrent neural networks, including GRU, BIGRU, LSTM, BILSTM, and Emotion–LSTM, across metrics such as accuracy, precision, recall, and F1 score. It is clear that Emotion–LSTM performs the best overall, indicating that the introduction of emotion vectors and their interaction with historical and current emotions is highly effective for music emotion classification tasks. Specifically, Emotio–LSTM exhibits higher accuracy, precision, recall, and F1 score in these experiments. This suggests that Emotion–LSTM can classify music emotions more accurately while maintaining a better balance, avoiding overfitting or underfitting issues.

Figure 7 provides a visual representation of the table's content, emphasizing the potential of the Emotion–LSTM model in the field of music emotion classification. It offers an effective approach to emotional analysis and lays a strong foundation for future research and applications. The introduction of emotion vectors and their interaction contributes to a more comprehensive and accurate understanding of the emotional elements present in music.

In Table 5, we also investigated the influence of the number of layers in the cross-processing module on the experimental results. We analyzed the experimental results on the music emotion dataset for cross-processing module layer numbers 1, 2, 3, 4, 5, and 6. From the experimental results, we can see that the performance does not linearly increase with the increase in the number of layers. The effectiveness of the model peaks when the number of layers in the cross-processing module is 3. Beyond 3 layers, both increasing and decreasing the number of layers leads to a corresponding decrease in experimental results. This may be due to the model's inability to capture the complexity of music emotion with too few layers, and the model becoming overly complex and prone to overfitting with too many layers. This finding

Table 4 Ablation comparison of different metrics and our model for different music levels in the music emotion recognition task

Model	DEAM dataset				FMA dataset			
	Accuracy %	Precision %	Recall %	*F*1 %	Accuracy %	Precision %	Recall %	*F*1 %
GRU	45.73	45.56	45.28	45.33	42.23	42.53	44.66	47.28
BIGRU	45.32	45.23	45.68	44.78	44.23	44.23	46.23	42.32
LSTM	47.53	47.23	47.52	47.23	47.52	46.23	46.33	46.23
BILSTM	48.23	47.25	47.66	46.78	47.53	47.23	46.23	46.23
Emotion–LSTM	49.68	50.06	50.17	49.84	49.54	49.42	49.55	49.89

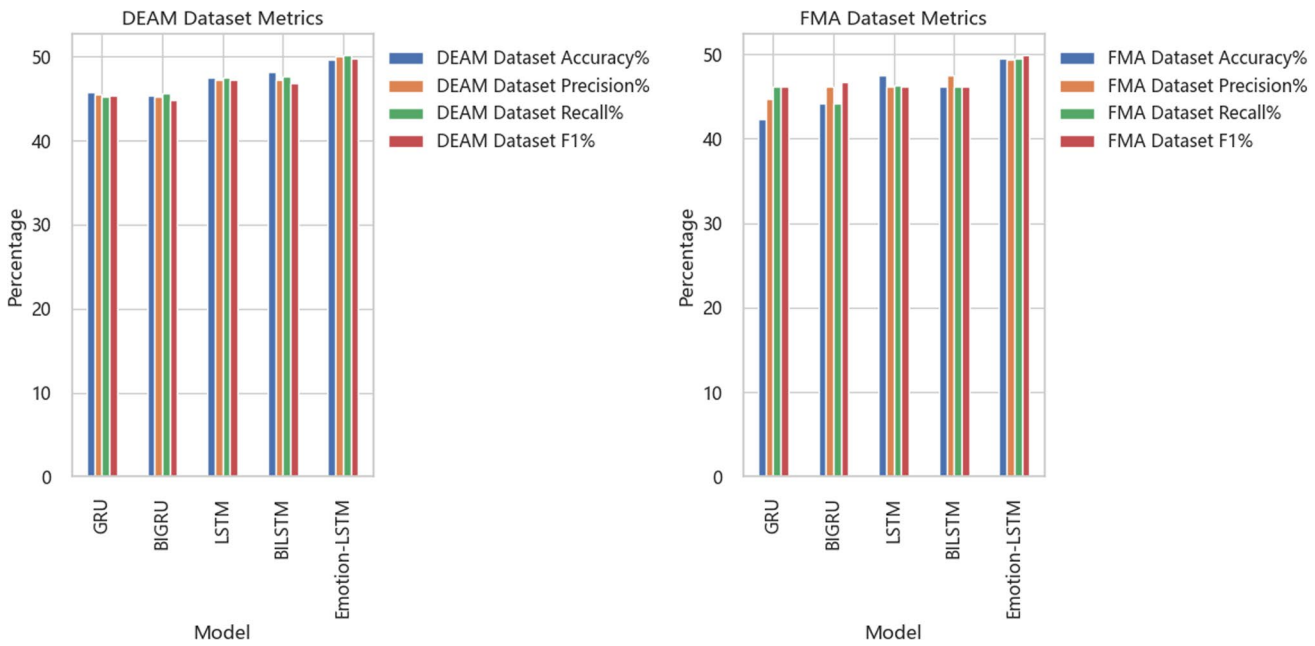

Fig. 7 Comparison of different indicators of different models

Table 5 The effect of the number of layers of the cross-processing module on the experimental results

	DEAM dataset				FMA dataset			
	Accuracy %	Precision %	Recall %	*F*1 %	Accuracy %	Precision %	Recall %	*F*1 %
1 layer	46.73	46.55	45.23	46.33	46.22	45.52	45.66	47.25
2 layer	47.53	47.23	47.52	47.23	47.52	46.23	46.33	46.23
`3 layer	49.68	50.06	50.17	49.84	49.54	49.42	49.55	49.89
4 layer	48.53	49.52	48.72	47.23	48.52	47.52	48.72	48.72
5 layer	48.52	48.52	46.32	47.53	48.52	47.36	48.33	48.25
6 layer	47.63	46.33	47.52	47.58	48.72	47.36	48.12	48.79

emphasizes the importance of carefully selecting the number of layers when designing deep learning models to maximize their performance. Furthermore, we also observed that the model's performance may have different effects on different music emotion tasks under different layer numbers. Therefore, in practical applications, researchers and practitioners need to balance and select the appropriate number of layers based on the specific task requirements and dataset characteristics to achieve the best performance and efficiency. Figure 8 visualizes the contents of the table.

6 Conclusion

In this study, we introduced the MMD-MII multimodal music emotion classification model, which integrates audio and lyric data while considering the inherent structure of music, including verses and choruses. Through experiments, we verified its effectiveness in capturing emotional information within music, showing promise for applications in music recommendation, advertising, psychology, and related fields.

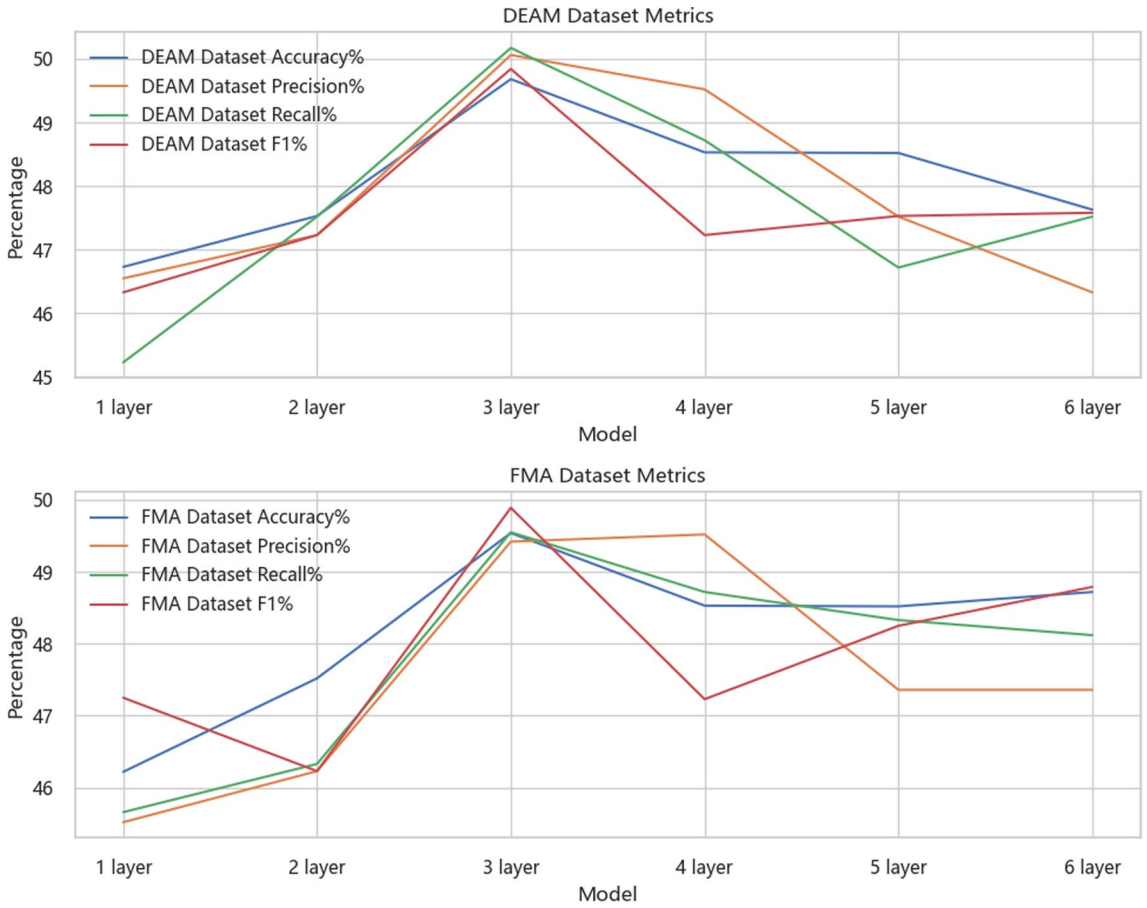

Fig. 8 Comparison of different indicators of different layer

However, our model still has limitations. Firstly, its performance may be constrained by the quality and diversity of input data. Additionally, despite incorporating music's inherent structure, there may be challenges in adapting to different music genres. These variations in emotional content and composition across genres can impact the model's accuracy and generalization. To address these challenges, future research will focus on enhancing the diversity and quality of training data and exploring techniques for genre-specific adaptation. We aim to improve the model's robustness and applicability across a broader range of music genres. Furthermore, we plan to expand its applications to real-world scenarios, including personalized music recommendations, advertising, and emotional therapy. Our research will continue to explore innovative methods and technologies to advance multimodal music emotion analysis.

In conclusion, the MMD-MII model represents a significant advancement in the field of music emotion classification. Despite the challenges and room for improvement, we believe that this research will provide valuable insights and methodologies for future multimodal music emotion analysis and related applications, ultimately contributing to a better music experience and emotional support for both individuals and society as a whole.

Author Contributions Jingyi Wang: Conceptualization, methodology, writing—review and editing. Alireza Sharifi: Formal analysis, visualization, software. Thippa Reddy Gadekallu: Supervision, methodology, writing—review and editing. Achyut Shankar: Methodology, writing—review and editing.

Funding This study did not receive any form of funding support.

Availability of Data and Materials Data and materials are not publicly available but can be requested from the author.

Declarations

Conflict of interest The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Pandeya, Y.R., Lee, J.: Deep learning-based late fusion of multimodal information for emotion classification of music video. Multimedia Tools Appl. **80**, 2887–2905 (2021)
2. Lucia-Mulas, M.J., Revuelta-Sanz, P., Ruiz-Mezcua, B., Gonzalez-Carrasco, I.: Automatic music emotion classification model for movie soundtrack subtitling based on neuroscientific premises. Appl. Intell. **53**, 27096–27109 (2023)
3. Hung, H., Ching, J., Doh, S., Kim, N., Nam, J., Yang, Y.: EMOPIA: a multi-modal pop piano dataset for emotion recognition and emotion-based music generation. arXiv preprint arXiv:2108.01374 (2021)
4. Chou, Y., Chen, I., Chang, C., Ching, J., Yang, Y., et al.: MidiBERT-piano: large-scale pre-training for symbolic music understanding. arXiv preprint arXiv:2107.05223 (2021)
5. Zheng, L.J., Mountstephens, J., Teo, J.: Four-class emotion classification in virtual reality using pupillometry. J. Big Data **7**, 1–9 (2020)
6. Jiang, D., Wu, K., Chen, D., Tu, G., Zhou, T., Garg, A., Gao, L.: A probability and integrated learning based classification algorithm for high-level human emotion recognition problems. Measurement **150**, 107049 (2020)
7. Sheykhivand, S., Mousavi, Z., Rezaii, T.Y., Farzamnia, A.: Recognizing emotions evoked by music using CNN-LSTM networks on EEG signals. IEEE Access **8**, 139332–139345 (2020)
8. Cunningham, S., Ridley, H., Weinel, J., Picking, R.: Supervised machine learning for audio emotion recognition: Enhancing film sound design using audio features, regression models and artificial neural networks. Pers. Ubiquit. Comput. **25**, 637–650 (2021)
9. Xing, B., Zhang, H., Zhang, K., Zhang, L., Wu, X., Shi, X., Yu, S., Zhang, S.: Exploiting EEG signals and audiovisual feature fusion for video emotion recognition. IEEE Access **7**, 59844–59861 (2019)
10. Wang, Z., Tong, Y., Heng, X.: Phase-locking value based graph convolutional neural networks for emotion recognition. IEEE Access **7**, 93711–93722 (2019)
11. Wu, S., Sun, M.: Exploring the efficacy of pre-trained checkpoints in text-to-music generation task. arXiv preprint arXiv:2211.11216 (2022)
12. Ocampo, R., Andres, J., Schmidt, A., Pegram, C., Shave, J., Hill, C., Wright, B., Bown, O.: Using GPT-3 to achieve semantically relevant data sonification for an art installation. In: International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar), , pp. 212–227. Springer (2023)
13. Chang, C., Lee, C., Yang, Y.: Variable-length music score infilling via XLNet and musically specialized positional encoding. arXiv preprint arXiv:2108.05064 (2021)
14. Alshanqiti, A., Namoun, A., Alsughayyir, A., Mashraqi, A.M., Gilal, A.R., Albouq, S.S.: Leveraging DistilBERT for summarizing Arabic text: an extractive dual-stage approach. IEEE Access **9**, 135594–135607 (2021)
15. Chen, H., Zhang, Z.: Hybrid neural network based on novel audio feature for vehicle type identification. Sci. Rep. **11**(1), 7648 (2021)
16. Mustaqeem, Kwon, S.: A CNN-assisted enhanced audio signal processing for speech emotion recognition. Sensors **20**(1), 183 (2019)
17. Wang, H., Gao, F., Zhao, Y., Wu, L.: WaveNet with cross-attention for audiovisual speech recognition. IEEE Access **8**, 169160–169168 (2020)
18. Shi, L., Du, K., Zhang, C., Ma, H., Yan, W.: Lung sound recognition algorithm based on VGGish-BiGRU. IEEE Access **7**, 139438–139449 (2019)

19. Zhang, Z., An, L., Cui, Z., Xu, A., Dong, T., Jiang, Y., Shi, J., Liu, X., Sun, X., Wang, M.: ABAW5 challenge: a facial affect recognition approach utilizing transformer encoder and audio-visual fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5724–5733 (2023)
20. Xu, S., Li, L., Yao, Y., Chen, Z., Wu, H., Lu, Q., Tong, H.: MUSE-NET: multi-scenario learning for repeat-aware personalized recommendation. In: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, pp. 517–525 (2023)
21. Zhu, R., Shi, L., Song, Y., Cai, Z.: Integrating gaze and mouse via joint cross-attention fusion net for students' activity recognition in e-learning. Proc ACM Interact Mob Wear Ubiquitous Technol **7**(3), 1–35 (2023)
22. Usmani, A., Alsamhi, S. H., Breslin, J., and Curry, E.: A novel framework for constructing multimodal knowledge graph from MuSe-CaR video reviews. In: 2023 IEEE 17th International Conference on Semantic Computing (ICSC), pp. 323–328 (2023)
23. Han, W., Jiang, T., Li, Y., Schuller, B., Ruan, H.: Ordinal learning for emotion recognition in customer service calls. In: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6494–6498 (2020)
24. Koh, E.Y., Cheuk, K.W., Heung, K.Y., Agres, K.R., Herremans, D.: MERP: a music dataset with emotion ratings and raters' profile information. Sensors **23**(1), 382 (2022)
25. Liu, K., DeMori, J., Abayomi, K.: Open set recognition for music genre classification. arXiv preprint arXiv:2209.07548 (2022)
26. Ding, Z., Qi, Y., Lin, D.: Albert-based sentiment analysis of movie review. In: 2021 4th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), pp. 1243–1246 (2021)
27. Kim, C. D., Kim, B., Lee, H., Kim, G.: AudioCaps: generating captions for audios in the wild. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 119–132 (2019)
28. Catharin, L. G., Ribeiro, R. P., Silla, C. N., Costa, Y. M. G., Feltrim, V. D. Multimodal classification of emotions in Latin music. In: 2020 IEEE International Symposium on Multimedia (ISM), pp. 173–180 (2020)
29. Pandeya, Y.R., Bhattarai, B., Lee, J.: Deep-learning-based multimodal emotion classification for music videos. Sensors **21**(14), 4927 (2021)
30. Zhao, J., Ru, G., Yu, Y., Wu, Y., Li, D., Li, W.: Multimodal music emotion recognition with hierarchical cross-modal attention network. In: 2022 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2022)
31. Chen, C., Li, Q.: A multimodal music emotion classification method based on multifeature combined network classifier. Math. Probl. Eng. **2020**(2020), 1–11 (2020)
32. Medina, Y.O., Beltrán, J.R., Baldassarri, S.: Emotional classification of music using neural networks with the MediaEval dataset. Pers. Ubiquitous Comput. **26**(4), 1237–1249 (2022)
33. Ning, E., Zhang, C., Wang, C., Ning, X., Chen, H., Bai, X.: Pedestrian Re-ID based on feature consistency and contrast enhancement. Displays **79**, 102467 (2023)
34. Wan, C., Wang, Y.: Node classification algorithm based on weighted meta-learning. J. Jilin Univ. Sci. Ed. **61**(2), 331–337 (2023)