Alexandria Engineering Journal 116 (2025) 586–600

Contents lists available at ScienceDirect

Alexandria Engineering Journal

journal homepage: www.elsevier.com/locate/aej

Original article

Real-time music emotion recognition based on multimodal fusion

Xingye Hao [a,b], Honghe Li [c,*], Yonggang Wen [d]

[a] Shanxi Shuozilun Symphony Orchestra, 030001, Shanxi, China
[b] College-Conservatory of Music, University of Cincinnati, 999039, Cincinnati, United States of America
[c] Humanities and Arts Media Department, Changzhi Medical College, 046000, Changzhi, China
[d] Department of Computer and Information Technology, Tianjin University of commerce Boustead College, 300384, Tianjin, China

ARTICLE INFO

Keywords:
Real-time emotion recognition
Multimodal fusion
Music emotion analysis
Feature fusion
Adaptive sampling
Affective computing

ABSTRACT

Multimodal emotion recognition is widely used in fields such as music emotion analysis and intelligent interaction. However, current models still face challenges in real-time and accuracy, especially in multimodal data fusion and emotional fluctuation processing. To this end, this paper proposes a real-time emotion recognition model based on Bi-LSTM and feature fusion, which effectively improves the capture efficiency of emotional features through multi-modal feature compression and adaptive sampling technology. The Bi-LSTM network is used to mine the time dependence of multi-modal data, while the feature fusion module integrates key emotional features in audio, visual and physiological signals, allowing the model to achieve a good balance between accuracy and real-time performance. Experimental results show that this model achieves higher accuracy and lower latency on the DEAP and AMIGOS data sets. Compared with existing methods, it has significant performance in multiple indicators such as weighted F1 score and G-Mean accuracy. To improve. Ablation experiments further confirmed that Bi-LSTM, feature fusion and adaptive sampling modules respectively make important contributions to the robustness and real-time performance of the model in emotion recognition tasks. This research provides an effective solution for multi-modal emotion recognition tasks, verifies the application potential of multi-modal feature fusion technology in music emotion analysis, and provides theoretical and practical support for the optimization of future emotional computing systems.

1. Introduction

Multimodal emotion recognition is a critical research area for intelligent human–computer interaction systems, aiming to capture the user's emotional state in real-time by integrating various modalities such as audio, visual, and physiological signals. This capability provides crucial support for personalized services and optimizing user experiences [1]. In the field of music emotion recognition, the integration of multimodal technologies offers rich emotional feedback for applications such as music recommendation, sentiment analysis, and virtual music assistants. This allows these systems to adapt more flexibly to user emotions and preferences. Emotion recognition technology has been gradually applied in diverse fields [2], including intelligent audio, virtual assistants, affective computing, and streaming media recommendation systems, significantly improving the adaptability of systems to users' emotional responses.

However, the real-time nature of emotion recognition presents significant challenges, particularly when processing multimodal data. Real-time recognition in high-dimensional, continuous, and dynamic data streams is often hindered by excessive computational overhead or insufficient response speed [3]. Optimizing the processing of these data streams while maintaining high recognition accuracy, in order to meet the requirements of low latency and high accuracy, has become a critical problem in the field of emotion recognition.

The advent of deep learning techniques has led to significant improvements in multimodal emotion recognition, particularly in the accuracy and feature extraction capabilities. Through the use of models like Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and autoencoders, emotion recognition systems can automatically extract and integrate high-dimensional features from audio, visual, and other data sources. This reduces the reliance on manual feature design and enhances the accuracy of emotion classification and prediction. In particular, deep learning techniques are effective in capturing the dynamic fluctuations in audio and video data, laying a solid foundation for emotion recognition in various contexts, including music emotion recognition, where deep learning models analyze the impact of music on user emotions.

Despite the impressive recognition accuracy achieved by deep learning, its computational cost remains a significant barrier to real-time

* Corresponding author.
E-mail addresses: 13453465296@163.com (X. Hao), lihonghe@czmc.edu.cn (H. Li), tjwyg_31@126.com (Y. Wen).

https://doi.org/10.1016/j.aej.2024.12.060
Received 17 November 2024; Received in revised form 11 December 2024; Accepted 16 December 2024
Available online 8 January 2025

applications. Most deep learning models rely on static architectures that are ill-suited for the dynamic nature of data streams. This results in models that struggle with recognition delays or degraded accuracy during periods of large emotional fluctuations. Furthermore, deep learning models often face issues with data redundancy and noise in multimodal data streams, making it difficult to balance computational efficiency with real-time performance. Therefore, the key challenge in emotion recognition lies in improving both the recognition accuracy and real-time processing performance, particularly in adapting to fluctuations in emotional data.

To address these challenges, this paper proposes a deep learning-based model for optimizing data flow in multimodal emotion recognition, enhancing both the real-time and adaptive nature of the system. The proposed model includes three key modules: a dynamic sampling and data cleaning module, a multimodal feature compression module, and an online adaptive emotion recognition module. First, the dynamic sampling and data cleaning module adjusts the sampling frequency based on emotional fluctuations, eliminating redundant information and noise, thus reducing the data stream burden while ensuring real-time data processing. Second, the multimodal feature compression module uses a pre-trained network to extract emotional features from audio and visual data and employs autoencoders to compress high-dimensional features, generating more compact data representations and reducing computational and storage costs. Finally, the online adaptive emotion recognition module employs transfer learning combined with small-batch incremental learning to enable the model to adjust its recognition capabilities in real-time, responding to dynamic changes in emotion data while maintaining high accuracy and stability.

The primary contribution of this paper is the proposal of a modular multimodal emotion recognition data flow optimization model, which enhances both real-time processing and adaptivity. The specific contributions are as follows:

(1) Proposing dynamic sampling mechanism based on emotion fluctuation: adjusting the frequency of data acquisition through dynamic sampling effectively reduces redundant information and lowers the load of data flow, which provides basic support for the real-time performance of the emotion recognition system.
(2) Designing an efficient compression module for multimodal features: feature compression of multimodal data using a self-encoder reduces computation and storage costs, enabling the model to operate efficiently even in resource-constrained environments.
(3) Introducing online adaptive learning strategy: small batch incremental learning is used to update the model parameters, which enhances the adaptivity of the model and enables it to maintain high recognition accuracy and robustness under the dynamic change of emotion data.

Through these modular designs, the proposed model not only achieves real-time emotion recognition in complex environments but also significantly improves the system's flexibility, providing a robust technical foundation for the practical application of multimodal emotion recognition systems.

The structure of this paper is organized as follows: Section 2 reviews the relevant research on multimodal emotion recognition and deep learning techniques for real-time data stream processing, analyzing the advantages and limitations of existing methods. Section 3 describes the algorithmic model proposed in this paper, detailing the design and implementation of the dynamic sampling and data cleaning module, the multimodal feature compression module, and the online adaptive emotion recognition module. Section 4 presents the experimental setup, evaluation metrics, and detailed analysis of experimental results, demonstrating the effectiveness of the model in terms of real-time performance and recognition accuracy. Finally, Section 5 summarizes the research findings, discusses the model's potential applications, and proposes future research directions.

2. Related work

Music-based multimodal emotion recognition technology has gradually become an important research direction in emotion computing and personalized recommendation systems. By fusing multimodal data such as audio, visual and physiological signals, music emotion recognition system can realize accurate capture of user's emotional state, thus supporting music recommendation, emotion regulation and adaptive interaction scenarios. However, as the demand for real-time interaction increases, the system needs to deal with high-dimensional, continuous, and dynamic data streams, and how to improve the data processing efficiency while ensuring the recognition accuracy has become a key challenge in this field. In this paper, we sort out related research from three aspects: multimodal music emotion recognition, real-time data stream processing, and online adaptive learning in deep learning, in order to better elucidate the research orientation of this paper.

2.1. Multimodal music emotion recognition

Multimodal music emotion recognition takes music audio as the core modality and combines it with other modal data such as visual or physiological signals to achieve more accurate emotion recognition. Unimodal music emotion recognition is usually based on audio features such as rhythm [4], frequency, pitch and other factors (e.g., the music emotion model proposed by Eerola and Vuoskoski, which predicts the emotional state through changes in the rhythm and pitch of the audio) [5], while the addition of visual and physiological signals provides more emotional features for the recognition, which makes the emotion recognition more comprehensive and accurate [4,6]. It has been shown that multimodal fusion in music emotion recognition can effectively enhance the robustness of the system [7]. For example, Koelstra et al. combined EEG (brainwave) signals and music audio information in the DEAP dataset, which significantly improved the accuracy of emotion recognition and the ability to capture emotional changes [8]. In addition, the dynamic changes in music in terms of emotion have strong temporal continuity and volatility, which makes how to effectively model audio emotion changes a core issue in multimodal music emotion recognition.

Currently, the research of multimodal music emotion recognition focuses on how to effectively integrate data such as audio, visual and physiological signals. Deep learning has made significant progress in feature extraction and modal fusion [9,10], and models such as convolutional neural networks (CNNs) and long-short-term memory networks (LSTMs) have demonstrated strong capabilities in modeling audio features and time-series data. The LSTM-based music emotion recognition model proposed by Neuman et al. accurate tracking [11, 12]. However, with the increase in the amount of emotion data, the application of traditional deep learning models in real-time recognition is limited [13], and the problem of high computational resource consumption is especially prominent in multimodal contexts. Therefore, how to realize efficient emotion recognition under multimodal data fusion has become a research focus in the field of music emotion recognition [14], and one of the core problems to be solved in this paper.

2.2. Real-time data stream processing technology

With the expansion of the application scenarios of the music emotion recognition system, the real-time demand of the system is increasing, and the real-time data stream processing technology has become a key link in emotion recognition [15]. In music emotion recognition, real-time data stream processing technology is mainly used to dynamically collect and process audio and other modal data to ensure the fast response of the emotion recognition system in user interaction [15,16]. Traditional music emotion recognition systems usually use a fixed sampling frequency for audio and visual data acquisition [17], but this

approach brings a large amount of redundant information and increases the computational and storage pressure of the system when the data changes frequently [18,19]. In recent years, researchers have proposed real-time data stream processing methods based on dynamic sampling to reduce the data stream burden by adjusting the sampling frequency. For example, abdellatif et al. proposed a dynamic sampling method based on music emotion fluctuation [20], which increases the sampling rate when the emotion changes are large and decreases the sampling frequency when the emotion changes are small to reduce redundant data.

Research hotspots in real-time data stream processing include techniques such as dynamic sampling, data cleansing and real-time compression [21], especially in music emotion recognition, where these methods can improve the data processing speed while ensuring the quality of the emotion recognition input [22]. Dynamic sampling techniques allow the system to adjust the sampling frequency in real time according to the emotion fluctuations in the audio, ensuring that only key data with significant emotion changes are retained [23]. Data cleaning techniques remove background noise and invalid information by filtering and de-noising, for example, through convolutional filters to clean the noise in audio and visual data and enhance the clarity of the data [24]. However, existing data stream processing methods mainly focus on data transmission and computational efficiency, and are less concerned with the accuracy enhancement of emotion recognition [25]. Therefore, how to optimize emotion recognition accuracy along with data stream processing remains an important challenge in the field of music emotion recognition.

2.3. Online adaptive learning in deep learning

The wide application of deep learning in music emotion recognition significantly improves the accuracy of emotion classification, especially in multimodal contexts to capture complex emotion changes [26]. However, traditional deep learning models usually adopt static architectures where the model parameters remain unchanged after training [27], making it difficult to adapt to the dynamic changes in the emotion data stream [28]. The dynamic nature of music emotions makes traditional static models unable to effectively track emotion fluctuations [29], especially in long time data streams, and the models show insufficient adaptability [30]. To solve this problem, online adaptive learning has been gradually applied to music emotion recognition, which allows the model to update its parameters according to new emotion data during the recognition process [31], thus improving the flexibility of recognition. Alshammar et al. proposed an online adaptive music emotion recognition model, which achieves gradual updating of the model through incremental learning [32], enabling it to adapt to real-time music emotion The model is gradually updated through incremental learning, so that it can adapt to the dynamic changes of music emotion in real time.

Currently, the application of online adaptive learning in music emotion recognition mainly focuses on how to balance the adaptability and stability of the model [33]. Incremental learning and migration learning are common online adaptive techniques, where incremental learning gradually updates model parameters with small batches of data [34,35], while migration learning uses pre-trained models to quickly adapt to new emotion data streams. However, the bottlenecks of online adaptive learning in real-time applications are high computational overhead and limited speed of model tuning [36]. As the data stream changes, the model needs to adapt quickly to avoid recognition delays or overfitting [37,38], which places high demands on computational resources and update efficiency. Therefore, designing efficient online adaptive learning mechanisms to ensure real-time and accurate music emotion recognition is a research hotspot in this field [39].

2.4. Limitations of existing methods

Although multimodal music emotion recognition, real-time data stream processing, and online adaptive learning play an important role in improving the real-time performance and accuracy of emotion recognition systems, the existing methods still face the following problems in practical applications: firstly, multimodal emotion recognition, despite its improved accuracy, results in high computational costs due to the high sampling frequency of audio and visual data in real-time music emotion recognition scenarios; secondly, the Real-time data stream processing methods mostly focus on reducing the transmission burden and improving the data quality, but pay insufficient attention to the final accuracy of emotion recognition; finally, although the online adaptive learning technique in deep learning improves the model's responsiveness to emotion fluctuations, it has a large computational overhead, and suffers from insufficient adaptation, especially in long-time dynamic data streams for music emotion recognition.

To cope with these problems, this paper proposes a multimodal music emotion recognition data stream optimization model that combines dynamic sampling, feature compression and online adaptation. The processing efficiency of the data stream is improved by dynamic sampling and data cleaning, and the parameter update of the model is realized by combining online adaptive learning to improve the accuracy of real-time music emotion recognition and the adaptability of the system.

3. Methodology

3.1. Dynamic sampling

In the real-time music emotion recognition system based on multimodal fusion, the dynamic sampling and data cleaning module is the key link, which is mainly responsible for reducing the unnecessary data burden, improving the real-time data processing efficiency, and at the same time guaranteeing the system's high quality of capturing emotion data. Multimodal data such as music audio and visual signals have continuous and high-dimensional characteristics, and the traditional fixed sampling rate scheme causes a large amount of data redundancy and increases the computational burden of transmission and processing. In the practical application of emotion recognition, this redundancy not only affects the real-time performance of the system, but also reduces the accuracy of capturing emotional states. The dynamic sampling and data cleaning module optimizes the system response speed and recognition effect by analyzing data fluctuations in real time, making adaptive adjustments to the sampling frequency, and improving the signal-to-noise ratio of the input data through noise cleaning.

In the design of this module, the dynamic sampling strategy is crucial. The core of this strategy lies in flexibly adjusting the sampling frequency according to the fluctuation of the emotion data to ensure the response and recognition efficiency of the system. First, in order to effectively analyze the emotional fluctuations in the audio data, the dynamic sampling module performs a Fast Fourier Transform (FFT) on the real-time sampled audio signals to convert the audio signals from the time domain to the frequency domain. This process is expressed by Eq:

$$X(f) = \sum_{n=0}^{N-1} x(n) \cdot e^{-j2\pi f \frac{n}{N}} \tag{1}$$

where $X(f)$ is the frequency domain representation of the audio signal, $x(n)$ is the time domain signal sampled at point n, N is the total number of samples, and f represents the frequency bin.

Based on the energy distribution of the frequency domain signal, the module further defines the formula for detecting emotion fluctuations, using the Frequency Variation Rate (FVR) to measure the energy changes in different frequency bands, in order to determine whether the sampling rate needs to be adjusted. Emotional fluctuations are usually

accompanied by significant changes in audio frequency, therefore, by detecting the fluctuation of energy at different frequencies, the key moments of emotional changes can be accurately identified, and the formula is as follows:

$$\mathrm{FVR} = \frac{1}{N_f} \sum_{f=1}^{N_f} |X(f) - X(f-1)| \tag{2}$$

where FVR is the frequency variation rate, N_f is the total number of frequency bins, $X(f)$ and $X(f-1)$ represent the current and previous frequency bin values.

In order to reduce the computational burden of the system, the dynamic sampling module dynamically adjusts the sampling rate according to the real-time value of the FVR. According to the current emotion fluctuation, the module sets the target sampling rate S_{target}, and the dynamic sampling rate S_{dyn} is adjusted by the following formula:

$$S_{\text{dyn}} = S_{\text{base}} \cdot (1 + \alpha \cdot \mathrm{FVR}) \tag{3}$$

where S_{dyn} is the dynamically adjusted sampling rate, S_{base} is the base sampling rate, and α is a scaling factor that controls the sensitivity of the sampling rate to the FVR.

When the emotion fluctuation is small and the data is stable, the module reduces the sampling rate to reduce redundant data, while the sampling rate is appropriately increased when the emotion changes intensely, so as to capture the emotion changes more accurately.

At the same time of data acquisition, the system also needs to perform data cleaning to remove noise and irrelevant information to ensure that the multimodal data input to the system has a high signal-to-noise ratio. In the music emotion recognition task, audio signals may contain background noise, and visual signals may be distorted due to changes in lighting or viewing angle, and these interfering factors can affect the accuracy of emotion recognition. For this reason, the data cleaning module gradually optimizes the audio and visual data to remove the background noise through filters and denoising algorithms. For example, in audio data, we use a combination of high-pass and low-pass filters to remove background noise at low or high frequencies to ensure data quality. The frequency response function of the bandpass filter is defined as follows:

$$H(f) = \begin{cases} 1, & f_{\text{low}} \le f \le f_{\text{high}} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $H(f)$ is the filter response at frequency f, f_{low} and f_{high} represent the lower and upper cutoff frequencies for the passband.

In this way, the system is able to remove data in irrelevant frequency bands, making the emotion features clearer and ensuring that the model focuses on key frequency bands when recognizing emotions.

In the data cleaning process, multimodal data also needs to be temporally aligned. In the multimodal emotion recognition system, modal data such as audio, visual and physiological signals are collected at different frequencies, and in order to ensure that different modal data are processed synchronously, the system uses a buffer or interpolation method for timing alignment. Assuming that the sampling interval of visual signals is T_v and the sampling interval of audio signals is T_a, the sampling point of the aligned data $y(t)$ can be calculated by the following formula:

$$y(t) = x_v\left(\frac{t}{T_v}\right) + x_a\left(\frac{t}{T_a}\right) \tag{5}$$

where $y(t)$ is the aligned signal at time t, $x_v\left(\frac{t}{T_v}\right)$ is the interpolated visual signal, and $x_a\left(\frac{t}{T_a}\right)$ is the interpolated audio signal.

Through timing alignment, the system is able to ensure that different modal data are synchronized within the same time window, providing high-quality data input for subsequent multimodal emotion feature extraction and recognition.

The output of the dynamic sampling and data cleaning module is a streamlined multimodal emotion data stream after sampling adjustment and noise removal. The output data not only reduces the overall data volume and reduces the computational pressure on the system, but also improves the signal-to-noise ratio of the data, thus enabling the emotion recognition model to achieve more accurate emotion state prediction based on high-quality data. The optimized data stream output from this module will be passed to the next module (feature compression module) to provide a reliable data base for real-time multimodal emotion recognition.

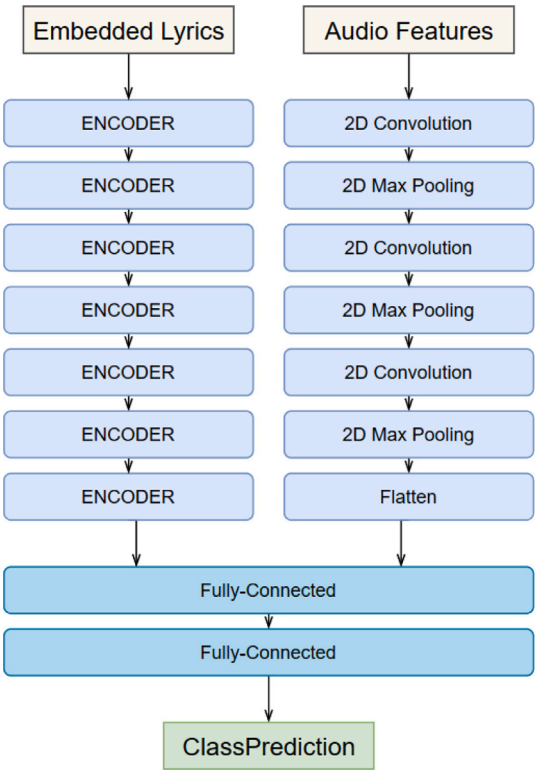

Fig. 1. Feature fusion structure diagram.

To create a multimodal representation, we fuse multiple features of the input, as shown in Fig. 1. Multimodal fusion is achieved by connecting the features generated by the model in the previous stage and inputting them into the fully connected layer for comprehensive processing. Multimodal fusion that combines audio and image information can generate richer feature expressions. This enhanced feature representation helps the model capture and understand the subtle differences in musical emotions more comprehensively. In this way, the model can use complementary information in different modalities to more accurately identify emotions.

In the diagram, the "Embedded Lyrics" section contains 7 encoders, each of which is responsible for gradually extracting the embedded features of the lyrics. These encoders are stacked in sequence, allowing the model to capture the semantic information of the lyrics from local to global. Specifically, these encoders generate high-dimensional representations with rich semantics by extracting features from the embedded lyrics layer by layer. This design ensures that the model has strong generalization ability in understanding the emotional expression and structural characteristics of the lyrics.

In the "Audio Features" section, the figure shows a feature extraction structure consisting of alternating 2D convolution layers and 2D max pooling layers. These layers extract important information in the spatial and frequency dimensions of the audio features by processing them. Finally, after flattening, a unified representation suitable for fusion with the lyrics features is generated.

By fusing the features of these two branches in the fully-connected layer, the system can combine the complementary characteristics of

lyrics and audio to build richer emotional expression features, thereby improving the accuracy and robustness of sentiment classification.

And in order to improve the real-time performance of the emotion recognition system, we introduced an adaptive sampling mechanism based on emotion fluctuation detection. This mechanism is designed to adjust the frequency of data acquisition dynamically in response to the emotional changes detected from the multimodal signals. The adaptive sampling process continuously monitors the fluctuation of emotional states through physiological and visual signals, and when emotional fluctuations are detected to be minimal, the system reduces the sampling rate to save computational resources. Conversely, when there are significant emotional shifts, the sampling rate is increased to capture the finer details of these changes, ensuring that the model maintains a high level of accuracy even during emotional transitions.

The effect of adaptive sampling on system latency is substantial. By reducing the sampling rate during periods of emotional stability, the system reduces the amount of data to process, thereby improving computational efficiency and reducing latency. This allows for faster response times and ensures that the model can operate in near real-time, which is crucial for practical applications such as real-time music emotion recognition. The dynamic adjustment of sampling frequency helps the system balance between maintaining high recognition accuracy and ensuring low latency.

3.2. Multimodal feature compression

In real-time music emotion recognition, due to the high-dimensional nature of multimodal data such as audio, visual, and physiological signals, the direct use of raw features for emotion recognition leads to excessive data processing overhead and tends to affect the real-time performance of the system. Therefore, the feature compression module reduces the computational cost and optimizes the system response speed by extracting and compressing core sentiment features from multimodal data to generate compact and efficient data representations. This module uses a pre-trained feature extraction network to extract key affective features from different modalities, and reduces the dimensionality of these high-dimensional features by means of a self-encoder to ensure that the data dimensionality is reduced while retaining as much affective information as possible.

Feature extraction is the first step in the multimodal feature compression module, aiming to extract core emotional features from audio and visual data. For audio data, we use a pre-trained convolutional neural network (CNN) model to extract emotionally relevant features such as tempo, frequency and pitch. For example, the audio signal $x_{\text{audio}}(t)$ is processed by the convolutional neural network to generate the feature matrix F_{audio}, which contains multi-level emotion information. Similarly, for visual data, features such as facial expressions and poses are extracted by pre-trained deep neural network models such as ResNet to generate visual feature matrices F_{visual}. These feature matrices will be used as input data for the subsequent compression process.

After completing the feature extraction, the module inputs the high-dimensional feature matrices into the auto-encoder for feature compression. Self-encoder is an unsupervised learning model that learns a low-dimensional representation of the data by encoding the input data into a low-dimensional space and decoding it back to the original data. Specifically, assume that the audio feature matrix is F_{audio} and the visual feature matrix is F_{visual}, and the compressed low-dimensional representations are notated as Z_{audio} and Z_{visual}. The encoding process of the self-encoder can be defined as:

$$Z_{\text{audio}} = f_{\text{encoder}}(F_{\text{audio}}) \tag{6}$$

where Z_{audio} is the compressed audio feature representation, and f_{encoder} represents the encoder function in the autoencoder model. Similarly, the compressed representation of the visual feature is:

$$Z_{\text{visual}} = f_{\text{encoder}}(F_{\text{visual}}) \tag{7}$$

where Z_{visual} is the compressed visual feature representation.

The coding process of the self-encoder is achieved by optimizing the reconstruction error. That is, during the compression and reconstruction process, the model retains the main information required for emotion recognition by minimizing the error between the input feature matrix F and the reconstruction matrix $\hat{F}$. The reconstruction error is defined as follows:

$$\mathcal{L}_{\text{reconstruction}} = \|F - \hat{F}\|_2^2 \tag{8}$$

where $\mathcal{L}_{\text{reconstruction}}$ is the reconstruction loss, F is the original feature matrix, and $\hat{F}$ is the reconstructed feature matrix after decoding. By minimizing this reconstruction loss, the self-encoder is able to retain key sentiment features with reduced data dimensionality.

In multimodal data, data features from different modalities will be further fused after compression to generate the final compressed feature representation. In order to realize the joint representation of multimodal features, the compressed features Z_{audio} and Z_{visual} will be concatenated or fused by a feature weighting strategy to generate the composite feature representation Z_{fusion}, with Eq:

$$Z_{\text{fusion}} = g_{\text{fusion}}(Z_{\text{audio}}, Z_{\text{visual}}, Z_{\text{physio}}) \tag{9}$$

where Z_{fusion} is the fused feature representation, Z_{audio}, Z_{visual}, and Z_{physio} are the compressed audio, visual, and the compressed audio, visual, and physiological feature representations respectively, and g_{fusion} represents the feature fusion function, which may involve concatenation of the feature fusion function. function, which may involve concatenation, weighted summation, or other fusion strategies.

Through the above process, the multimodal feature compression module is able to effectively reduce the data dimensions and output compact and efficient emotion feature representations for real-time music emotion recognition. This module not only significantly reduces the transmission and storage pressure of the data stream, but also ensures that the compressed feature data can still accurately reflect the user's emotional state by maintaining the integrity of the emotional features. The final compressed features will be passed to the emotion recognition module for real-time emotion classification and prediction, thus realizing efficient multimodal emotion recognition.

3.3. Online adaptive emotion recognition

In the real-time music emotion recognition system, the traditional static emotion recognition model is difficult to maintain high accuracy and robustness over a long period of time because the user's emotional state and multimodal data streams change dynamically over time. For this reason, the online adaptive emotion recognition module adopts an incremental learning strategy combined with a bidirectional long-short-term memory network (Bi-LSTM) structure, so that the model can adjust its parameters over time and gradually adapt to the changes of multimodal emotion features in the real-time data stream. Through the design of this module, the system is able to dynamically respond to emotion fluctuations and data changes, effectively improving the real-time and accuracy of recognition.

Bi-directional long and short-term memory network (Bi-LSTM) has a significant advantage in capturing time series features, and is able to extract more comprehensive emotion information in the front and back time steps of the time series. This is especially important for music emotion recognition, as music data itself has strong time-dependence and emotion fluctuation, e.g., changes in tempo and pitch can cause fluctuations in the user's emotional state. Bi-LSTM is able to capture emotion features from both past and future time steps through the bi-directional processing mechanism, which realizes a more accurate description of the emotional state. This bi-directional processing is particularly effective in capturing the subtle changes in music emotion over time, allowing the system to detect changes in the user's emotional response in real time as the music emotion produces fluctuations. Bi-LSTM has forward and backward processing mechanisms, which can

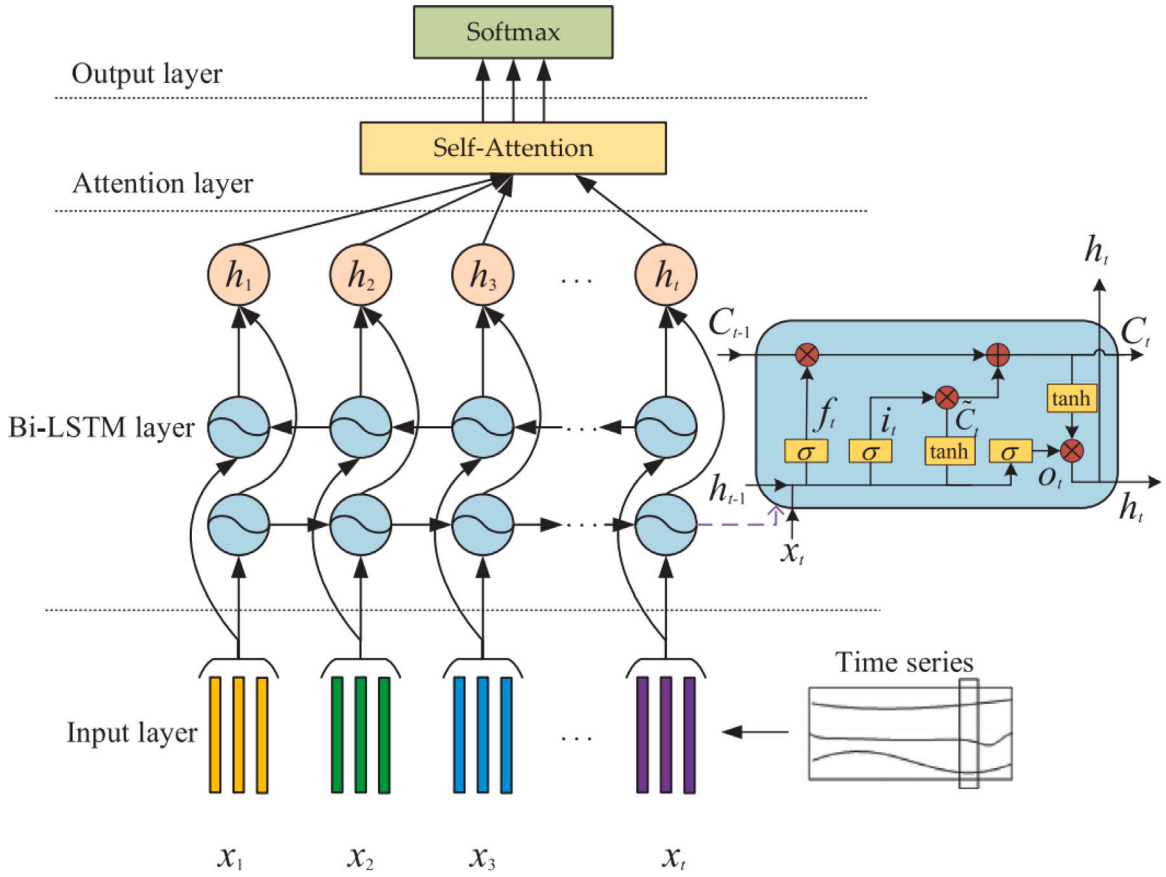

Fig. 2. Bi-LSTM architecture for temporal feature extraction in emotion recognition.

focus on the past and future of the time series at the same time, thereby extracting more comprehensive emotional features. As shown in Fig. 2, the Bi-LSTM structure shows how to transfer information between different time steps, using the bidirectional feature capture capability to effectively identify emotional changes, providing strong support for multimodal emotional fusion.

The main use of this module is to adapt to emotional changes by updating model parameters based on real-time input of multimodal data. Specifically, the online adaptive emotion recognition module gradually adjusts the weights of the model through incremental learning, so that the Bi-LSTM model can be continuously updated in small increments to adapt to the user's current emotional state when new emotional data are input. Assuming that the input of the model is a comprehensive feature representation Z_{fusion}, the output predicted value of emotion recognition can be expressed as:

$$\hat{y}_t = f_{\text{emotion}}(Z_{\text{fusion}}; \theta_t) \tag{10}$$

where $\hat{y}_t$ is the predicted value of the emotion state at time step t, f_{emotion} denotes the emotion recognition function, and θ_t is the current parameters of the model. After each new data input, the model updates the parameters through incremental learning of small batches of data, so that the parameters θ gradually adapt to the new sentiment data while retaining the existing information.

On the basis of incremental updating, the online adaptive module also adopts a lightweight migration learning strategy. When there is a significant change in the sentiment data, the system uses migration learning to quickly adjust the model weights so that the model can quickly adapt to the new sentiment context. By referring to the weights of the pre-trained model, this approach enables the model to transition more smoothly in the face of dramatic emotional fluctuations and reduces the computational burden caused by frequent updates. Overall, this adaptive updating mechanism based on Bi-LSTM enables the model to not only maintain high recognition performance in emotion data for a long period of time, but also quickly adjust to ensure real-time performance when encountering new emotion states.

To ensure the good performance of the Bi-LSTM model in the emotion recognition task, we carefully tuned its hyperparameters, especially the batch size and learning rate. The following is a detailed explanation of the selection of these hyperparameters:

- Batch Size: The batch size determines the number of samples used for each parameter update. In our experiment, the batch size was selected as 32. A smaller batch size can update the model parameters more frequently, which helps to accelerate convergence. At the same time, a batch size of 32 can also ensure efficient training under limited computing resources and avoid insufficient memory due to too large batches. Through experimental tuning, we found that when the batch size is 32, the training effect and computing efficiency are well balanced.
- Learning Rate: The learning rate controls the step size of each parameter update. In our experiment, we chose an initial learning rate of 0.001, which is an empirical value commonly used in the field of deep learning. This learning rate setting can maintain a stable convergence speed during a long training process. In order to further improve the training efficiency and prevent overfitting in the later stages of training, we also adopted a learning rate decay strategy to gradually reduce the learning rate during training, thereby ensuring that the model is more finely tuned in the later stages.
- Other hyperparameter settings: In addition to batch size and learning rate, we also adjusted other hyperparameters. The number of LSTM units was set to 128 to ensure that the model had sufficient expressive power but did not cause overcomputation due to too many units. To prevent overfitting, the dropout rate

was set to 0.5, which helped to enhance the generalization ability of the model.

Through the optimization of these parameters, the Bi-LSTM model showed good real-time performance and accuracy in the multimodal emotion recognition task.

This module has an important role in the overall model of this paper. First, it is an important module to ensure the accuracy of emotion recognition. Through the incremental learning method of real-time updating, the system can adjust the model parameters in time when the user's emotional state changes, making the recognition results more accurate. Second, this module effectively improves the robustness of emotion recognition, especially when dealing with long time and frequent emotion fluctuation music situations, the bi-directional feature of Bi-LSTM is able to capture more complete emotion information, ensuring that the system is able to make a smooth transition in the user experience. Third, the online adaptive emotion recognition module provides a guarantee for the real-time performance of the system, which enables it to flexibly cope with a variety of complex emotional fluctuations in practical applications and provides reliable technical support for the stability and adaptability of music emotion recognition.

The online adaptive emotion recognition module provides core support for real-time music emotion recognition by combining Bi-LSTM and incremental learning to ensure that the system can dynamically adjust to changing emotion data, balancing the needs for accuracy and real-time performance. The design of this module not only realizes efficient emotion recognition, but also enhances the model's ability to adapt to fluctuations in emotion data, making the system smarter and more flexible in the practical application of music emotion recognition.

4. Experiment

4.1. Datasets

In order to evaluate the effectiveness and accuracy of the real-time multimodal music emotion recognition model proposed in this paper, the experiment uses two well-established and authoritative emotion datasets in the field of multimodal emotion recognition that contain music-related data. The selected datasets, DEAP and AMIGOS, cover multimodal data such as audio, visual, and physiological signals, which can reflect the diversity and continuity of emotional changes and provide comprehensive testing conditions for complex situations in emotion recognition tasks. These datasets were chosen for their relevance to real-time emotion recognition in music-based applications, as they contain diverse multimodal features and rich emotional labels that allow for a comprehensive evaluation of the proposed model's performance. Below is a detailed description of these two datasets and their relevance to the study.

The DEAP (Dataset for Emotion Analysis using Physiological signals) Dataset [40] is one of the standard datasets widely used in emotion recognition research and is specifically designed for multimodal emotion recognition tasks. It contains physiological signals and facial video data collected from 32 participants during the viewing of 40 music video clips, with the aim of recording the emotional responses generated by users while viewing these clips. The DEAP dataset provides rich physiological signals, including electroencephalogram (EEG), galvanic skin conductance (GSR), electrocardiogram (ECG), as well as facial video data. This makes it an ideal dataset for testing emotion recognition systems that integrate both physiological and visual data. The DEAP dataset is particularly valuable for real-time emotion recognition because it captures the continuous emotional impact of music on users, with each clip lasting around 60 s and labeled across three emotional dimensions: Valence, Arousal, and Dominance, on a scale from 1 to 9. In this study, the high-frequency physiological signals were downsampled to improve the real-time performance and computational efficiency of the model, making it suitable for real-time applications. The multimodal nature of the DEAP dataset, along with its comprehensive emotion labels, is crucial for evaluating real-time emotion recognition systems, particularly in the context of music and emotional responses.

The AMIGOS (A Dataset for Affect, Personality and Mood Research on Individuals and Groups) Dataset [41] is another highly relevant dataset for music emotion recognition tasks. The AMIGOS dataset also contains multimodal data, including audio signals, visual signals (face video), and physiological signals (EEG, GSR, ECG), collected during the viewing of music videos. This dataset focuses specifically on the emotional responses evoked by music, making it particularly suited for our study, which aims to analyze the influence of music on emotions. The AMIGOS dataset is labeled with emotional dimensions (Valence, Arousal) and emotion polarity labels (Positive/Negative) for each segment, which allows us to comprehensively assess the emotional fluctuations brought on by different music videos. The dataset has a sampling rate of 128 Hz for physiological signals and 30 frames per second for facial video, which aligns well with real-time emotion recognition tasks, where low-latency processing and multimodal fusion are critical. The segmented structure and rich emotion labels of AMIGOS allow for an in-depth evaluation of the proposed model's performance in diverse emotional contexts, making it an important dataset for testing real-time emotion recognition systems in music-related applications.

Both the DEAP and AMIGOS datasets are highly relevant for this study due to their multimodal nature and focus on music-induced emotions. The inclusion of both physiological signals and facial expressions enables a robust evaluation of the model's ability to integrate diverse emotional features. By using these datasets, the experiment is able to comprehensively assess the performance and adaptability of the proposed real-time multimodal emotion recognition model, ensuring that it is capable of handling complex, dynamic emotional data in practical applications.

To ensure the efficiency and accuracy of the model, we preprocessed the data before the experiment. The specific preprocessing steps include the following: First, the high-frequency physiological signal data were downsampled to reduce the data transmission and processing overhead. The EEG data in DEAP was downsampled from 512 Hz to 128 Hz, while the physiological signal in AMIGOS was downsampled to 64 Hz. Additionally, to remove low-frequency and high-frequency noises in the physiological signals, we applied band-pass filters to retain the relevant emotional features and discard irrelevant noise.

For visual data, we applied noise reduction techniques to improve the quality of the facial video signals. Specifically, we used background subtraction algorithms to remove static background noise from the videos, ensuring that only the face area was focused on for emotion recognition. To further enhance data quality, light normalization techniques were applied to account for varying lighting conditions, which helped reduce the impact of illumination changes on emotion detection.

For audio data, noise reduction was crucial for removing environmental interference. We applied spectral subtraction techniques to clean the audio signals and reduce background noise. Additionally, the audio was normalized to ensure consistent volume levels across different samples, which minimized the potential for audio volume-related bias during emotion recognition. We also removed periods of silence or low energy in the audio signals, focusing the model's attention on the more informative sections of the music. Finally, we performed temporal alignment of audio, visual, and physiological signal data to ensure the synchronization of each modality data within the same time window, which facilitates multimodal feature fusion to provide a unified data base.

The important contribution of DEAP and AMIGOS datasets in this experiment is to provide a high-quality data source for multimodal emotion recognition. Both of them not only contain rich emotion labels and multimodal features, but also test the generalization ability of emotion recognition models through the design of different contexts. With these two datasets, this experiment is able to comprehensively verify the effectiveness and adaptability of the real-time multimodal emotion recognition model proposed in this paper under different contexts and emotion fluctuations, which provides reliable experimental support for the model's performance in practical applications.

Table 1
Experimental environment configuration.

Component	Specification
Processor (CPU)	Intel Core i7-8700K, 3.7 GHz
Graphics Card (GPU)	NVIDIA GeForce GTX 1080 Ti, 11 GB GDDR5X
Memory (RAM)	32 GB DDR4
Storage	1 TB SSD
Operating System	Ubuntu 20.04 LTS
Programming Language	Python 3.8
Deep Learning Framework	TensorFlow 2.4
CUDA Version	CUDA 11.1
CuDNN Version	CuDNN 8.0

4.2. Details

In order to validate the performance of the real-time music emotion recognition model based on multimodal fusion proposed in this paper, the experiments are comprehensively designed from data processing, model training to evaluation to ensure that the recognition accuracy, real-time performance and robustness of the model are fully examined in different emotion contexts. The experimental environment, data preprocessing, model training process and key settings all provide comprehensive support for the effectiveness of the model in this paper.

The experiments were conducted on a high-performance workstation configured with NVIDIA GTX 1080 Ti GPUs with Intel i7 CPUs and 32 GB of RAM, and the experiments were implemented using the Python programming language and based on the TensorFlow deep learning framework in order to make full use of the computational power of the GPUs to accelerate the model training and inference process. In the experimental environment, relying on the advantages of GPU acceleration, it is possible to realize the real-time processing of multimodal data under the premise of guaranteeing the model complexity, so as to support the fast computation of large-scale data in emotion recognition tasks. This Table 1 below contains more detailed information on the experimental environment available.

The experimental dataset adopts the DEAP and AMIGOS datasets, both of which contain multimodal data such as audio, visual and physiological signals. Prior to the experiments, several preprocesses were performed on the datasets to enhance the computational efficiency and recognition performance of the models. The specific preprocessing steps include the following aspects:

1. Preprocessing of physiological signals
   Physiological signals include electroencephalogram (EEG), galvanic skin response (GSR) and electrocardiogram (ECG), which have high-frequency noise and low-frequency fluctuations. Therefore, in the preprocessing process, we first performed filtering operations to remove low-frequency and high-frequency noise to ensure the purity of the signal. Specifically, for EEG signals, we use band-pass filters to remove noise below 0.5 Hz and above 50 Hz. Since the dimensions of different physiological signals vary greatly, we also normalize them. The normalization method used is Z-score standardization, that is, for each signal feature, its mean is set to 0 and the standard deviation is set to 1, so that the values of various signals can be compressed to the same scale to avoid model deviations caused by different dimensions.
   For the EEG signals in the DEAP dataset, we reduced the sampling rate from 512 Hz to 128 Hz to reduce the calculation amount of high-frequency signals. In the AMIGOS dataset, the sampling rate of physiological signals was adjusted to 64 Hz.
2. Identify and eliminate emotional noise
   In this study, we applied effective identification and removal techniques for emotional noise in multimodal data. Audio data was removed from background noise and distortion using low-pass and high-pass filters and spectral subtraction; visual data was removed using facial feature point detection, background removal, and motion tracking techniques to ensure that only facial information related to emotion was retained; noise in physiological signals was removed using bandpass filters and independent component analysis (ICA) algorithms. To further improve data quality, we also combined adaptive sampling technology to identify and filter irrelevant emotional fluctuations in real time to ensure accurate capture of emotional features. These noise removal methods effectively improved the quality of multimodal data and improved the accuracy and real-time performance of the emotion recognition system.
3. Preprocessing of audio data
   Audio data is another important modality of this study. First, we adjusted the sampling rate of the audio signal to 16 kHz, and denoised the audio signal to remove background noise and ensure the quality of the audio data. The preprocessing step of the audio signal also includes short-time Fourier transform (STFT), which extracts frequency domain features and provides strong support for subsequent emotion analysis.
   In order to be consistent with physiological signals and facial video data, we normalized the audio signal and used Min–Max Normalization to compress each feature value of the audio signal to the [0, 1] interval. This normalization method helps to eliminate the dimensional differences between different audio features and avoid the unbalanced impact of audio features on the model during training.
4. Multimodal Data Alignment
   Due to the different sampling frequencies of audio, visual, and physiological signal data, we time-aligned all signals before multimodal data fusion. Specifically, we first determined a unified time window to synchronize the audio, facial video, and physiological signal data to the same time axis for subsequent feature extraction and fusion.

During the model training process, the experiment first inputs the pre-processed audio, visual and physiological signal data into the dynamic sampling and data cleaning module to further optimize the quality of the data and computational efficiency. This module effectively reduces the data volume under the premise of ensuring the integrity of the emotion data, enabling the system to process multimodal emotion data more efficiently. Next, the data stream enters the multimodal feature compression module, which generates compact multimodal feature representations by performing dimensionality reduction operations on the audio and visual feature matrices through the self-encoder. This process ensures that the key information of the emotion features is preserved in the low-dimensional space after data compression, providing a lightweight and efficient data input for real-time emotion recognition.

In the model training phase, the model in this paper mainly relies on a bidirectional long-short-term memory network (Bi-LSTM) to capture the temporal dependence in the multimodal affective features. The bidirectional structure of the Bi-LSTM model is able to efficiently deal with complex affective fluctuations in both audio and physiological data, allowing it to be responsive to temporal changes in the affective state. The model training uses the Adam optimizer to optimize the parameters of Bi-LSTM, and the loss function is cross-entropy loss. In order to enhance the adaptivity of the model, this paper introduces an online incremental learning strategy in order to gradually update the Bi-LSTM model parameters during the training phase. The goal of incremental learning is to adjust the model parameters to maintain a flexible response to changes in the sentiment data through step-by-step training with small batches of data. The training process adopts an 8:2 dataset division, i.e., 80% of the data is used for model training and 20% is used for testing and validation to ensure the model's generalization ability during the training process.

During the training process, each iteration of the model is optimized on different emotional contexts and multimodal feature combinations,

Table 2
Linking structural elements of music to emotional responses.

Musical feature	Description	Associated emotions
Tempo	Speed or pace of music	Fast: excitement, happiness, anger; Slow: calm, sadness, tranquility
Mode	Type of musical scale	Major: joy, optimism; Minor: sorrow, introspection
Loudness	Intensity and amplitude of sound	High: power, aggression; Low: peace, relaxation
Pitch	Perceived frequency of a sound	High: excitement, tension; Low: melancholy, calm
Melody	Sequence of notes perceived as a cohesive whole	Harmonious: joy, relaxation; Dissonant: unease, tension
Rhythm	Pattern of beats and timing	Steady: peace, happiness; Irregular: surprise, tension
Harmony	Combination of notes or chords	Consonant: comfort, joy; Dissonant: anxiety, suspense
Timbre	Tone color or quality of sound	Bright: cheerfulness; Dark: sadness, mystery
Dynamics	Variation in loudness across a piece	Crescendo: excitement; Decrescendo: calmness
Articulation	Style of note execution (e.g., staccato, legato)	Staccato: playfulness, energy; Legato: serenity, smoothness
Contour	Shape of pitch movement over time	Ascending: hope, excitement; Descending: sadness
Texture	Layers and interaction of musical lines	Thick: intensity, power; Thin: simplicity, loneliness

and the model training process records the training loss and accuracy changes to ensure that the model converges to a stable recognition effect. In addition, in order to improve the model's ability to respond quickly to emotion fluctuations, this paper designs a transfer learning strategy based on emotion fluctuation detection. When the system detects a dramatic change in the emotion state, the migration learning mechanism loads the pre-trained weights to quickly adapt to the new emotion state, thus reducing the computational cost required for model adjustment. With this strategy, the Bi-LSTM model is able to maintain high robustness and real-time performance in different emotional contexts.

Through this experimental setup and training process, the model is able to optimize the real-time performance while maintaining the recognition accuracy, which provides a technical guarantee for the practical application of music emotion recognition systems. The efficiency of the experimental environment, the refinement of the preprocessing, and the design of the training process all provide comprehensive support for the performance of the model in multimodal emotion recognition in this paper.

4.3. Results

Numerous studies have demonstrated that a range of audio characteristics significantly contribute to the analysis of emotion in music. These investigations examined how people emotionally respond to musical compositions and identified specific audio features that are tightly linked with emotional reactions. Key elements include pitch, loudness, audio intensity, and tempo variations. Through examining these aspects, researchers can detect patterns that align closely with emotional expression. For instance, shifts in pitch may express sadness or joy, whereas loudness and energy can evoke powerful or subdued emotional experiences. Additionally, tempo and speed modifications influence the emotional momentum and rhythm within a musical piece. Analyzing these features computationally has provided researchers with a profound insight into the subtleties of emotion conveyed in music. Table 2 presents typical music genres alongside their associated emotional classifications.

To analyze audio efficiently and extract emotion-related information, we employ diverse technical methods from digital signal processing. These methods utilize sophisticated algorithms capable of identifying emotional characteristics within audio signals. By employing spectral, time-domain, and time–frequency analyses, essential details regarding emotion-based features within audio can be accurately captured.

The Short Time Fourier Transform (STFT) generates a time-frequency representation of the audio "spectrogram" that reveals the frequency components of the audio signal over time. To calculate the STFT, the signal is divided into segments of equal length, and the Fourier Transform is then applied to each segment, producing a spectrogram that shows frequency information across time. In the case of continuous time, the function that is desired to be transformed is denoted as $x(t)$, and the convolution operation is performed by means of a short-time activated window function $w(t)$. The formula is as follows:

$$STFT\{x(t)\} \equiv X(\tau, w) = \int_{-\infty}^{+\infty} x(t)\, w(t-\tau)\, e^{-it} dt \tag{11}$$

Mel Spectrogram (Mel Spectrogram) in the traditional spectrogram based on the frequency converted to mimic the human ear perception of the Mel scale (Mel scale), more closely related to the human auditory system of tone perception, the calculation formula is as follows:

$$mel(f) = \frac{1000}{\log_{10} 2} \log_{10}\left(1 + \frac{f}{1000}\right) \tag{12}$$

On the basis of Mel spectrum, logarithmic transformation is applied to obtain Log-Mel Spectrogram, and then Mel Frequency Cepstrum Coefficient (MFCC) can be obtained by linear cosine transformation. The formula is:

$$MFCC = \sqrt{\frac{2}{M}} \sum_{M=1}^{M} X_m(i) \cos\left(\frac{c\pi\left(m - \frac{1}{2}\right)}{M_m}\right) \tag{13}$$

Chroma features play a vital role in music information retrieval as they capture essential information linked to harmony. These features remain stable despite variations in timbre and directly correspond to musical harmony. According to Müller, chroma features are particularly effective for extracting mid-level audio characteristics. Assuming a Western scale, the vector x is derived by calculating the spectrogram of the signal within each time window, with components $x = [x_1, x_2, \ldots, x_{12}]$ representing the frequencies of each pitch class.

$$\{C, C^{\#}, D, D^{\#}, E, F, F^{\#}, G, G^{\#}, A, A^{\#}, B\} \tag{14}$$

In this experiment, we extracted these audio features for further testing, exploring various combinations to identify those most informative for emotion recognition. This approach enabled us to pinpoint emotion-related features essential to the task, refining the feature set through iterative experimentation. Next, we trained a GRU model on audio data sequences, using the selected feature data to detect emotional content in the music.

The following images (Figs. 3, 4) illustrate the extracted features of a sampled track from the dataset, namely "Matsuri" by Fuji Kaze. This track falls under the genre of contemporary Japanese pop, characterized by a vibrant blend of instrumental and vocal elements. Its emotional label is identified as "energetic", fitting its upbeat tempo and dynamic rhythmic patterns. The song was analyzed in a compressed wav file format to capture the intricate layers of sound production, including changes in melody, tempo, and timbre that contribute to its emotional expression. Given our study's focus on real-time multimodal emotion recognition, this sample offers insights into how auditory features like pitch, loudness, and rhythm interact to convey high-energy emotions. These extracted features serve as foundational data for our model, showcasing the diverse musical elements that influence emotional perception in real-time analysis.

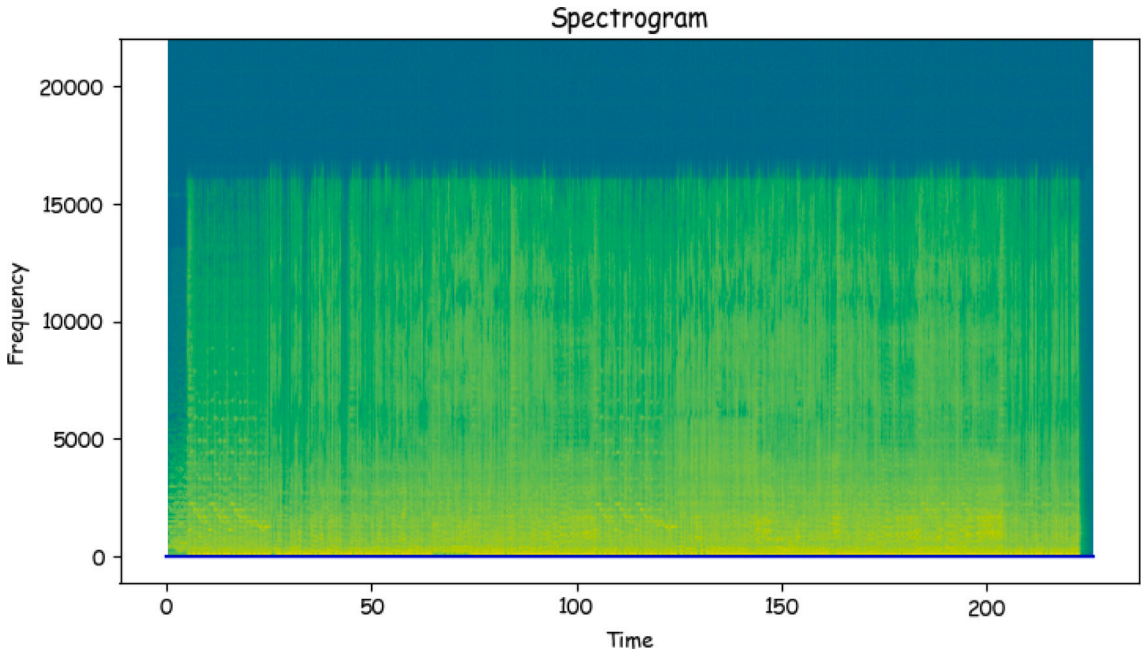

Fig. 3. Spectrogram of a musical track.

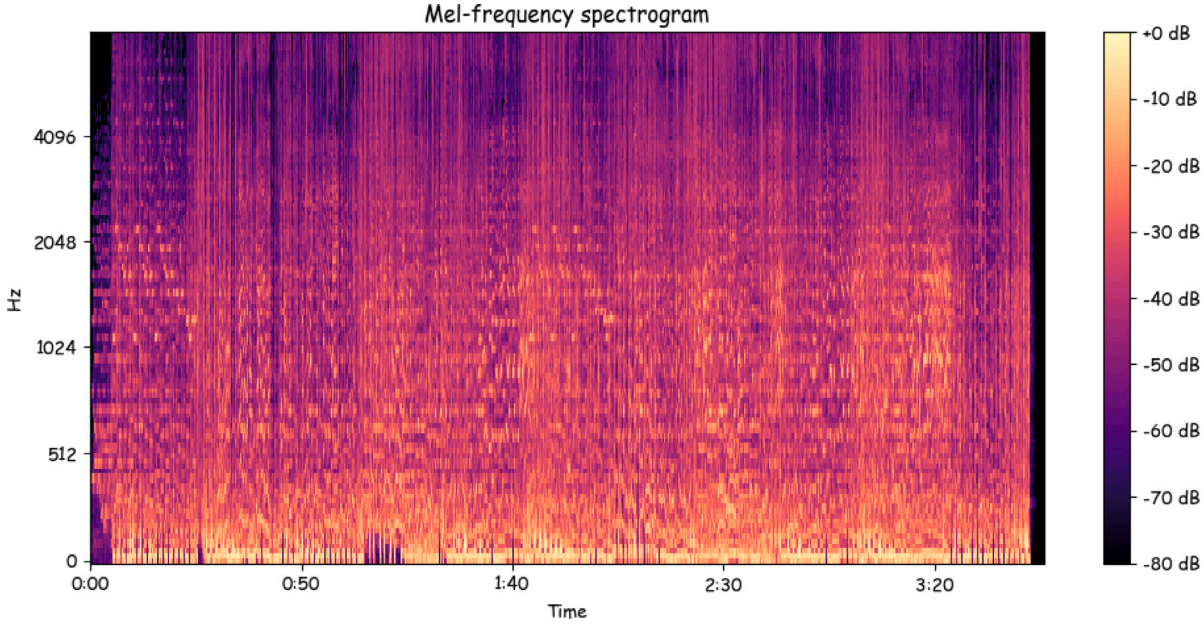

Fig. 4. Mel Spectrogram of a musical track.

In the emotion recognition experiments, in order to comprehensively validate the performance advantages of our proposed model based on Bi-LSTM and feature fusion on the DEAP dataset, this paper demonstrates a detailed comparison between our model and multiple other benchmark models in terms of accuracy, F1 score, precision, recall and latency in Table 3. This table reveals the specific performance of each model in multimodal emotion recognition tasks, reflecting the significant advantages of our approach in various aspects.

As can be seen from the table, our model outperforms other models on several key metrics. In terms of accuracy, our model achieves 87.5%, which is a 4.3 percentage point improvement over the classic CNN-RNN benchmark model (83.2%), and even has a significant advantage when comparing it to the LSTM + Attention model (85.3%), which has a higher modeling capability. This advantage mainly comes from the Bi-LSTM structure we adopt, which is able to bi-directionally capture the emotion fluctuation features of the time series, and combined with the feature fusion strategy, it enables the multimodal features of audio, visual and physiological signals to be more adequately fused in the time-series and feature space. In addition, the dynamic update mechanism based on Bi-LSTM allows the model to maintain an efficient response under different emotional states, avoiding the problem that traditional models are difficult to cope with emotional changes.

In terms of F1 score, our model achieves 0.89, which is 6 and 7 percentage points higher than the 0.83 and 0.82 of EmotionNet (CNN) and Hybrid CNN-GRU models, respectively. This indicates that our method not only excels in prediction accuracy, but also is more robust and adaptable in handling emotion recognition tasks with unbalanced data distribution. Further analyzing the precision and recall, our model

Table 3
Comparison of our model with other state-of-the-art models on the DEAP dataset.

Model	Accuracy (%)	F1 score	Precision	Recall	Latency (ms)	Average score
Our model (Bi-LSTM + Feature fusion)	**87.5**	**0.89**	**0.88**	**0.87**	**24.3**	**0.88**
CNN-RNN (Baseline) [42]	83.2	0.82	0.81	0.82	30.5	0.82
ST-GCN [43]	79.6	0.79	0.78	0.79	35.8	0.79
MLP-EmotionNet	81.3	0.81	0.80	0.81	28.4	0.80
SVM + Feature extraction [44]	75.8	0.74	0.75	0.73	41.2	0.74
GRU + Spectrogram features [45]	78.9	0.77	0.78	0.77	29.1	0.77
EmotionNet (CNN) [46]	84.1	0.83	0.82	0.83	26.7	0.83
Bi-GRU [47]	82.4	0.81	0.80	0.81	27.9	0.81
LSTM + Attention [48]	85.3	0.85	0.84	0.85	25.6	0.85
Capsule network [49]	80.1	0.79	0.78	0.79	33.4	0.80
Deep belief network [50]	76.5	0.75	0.75	0.74	37.5	0.75
Hybrid CNN-GRU [51]	83.7	0.82	0.83	0.82	28.1	0.82
Random forest classifier [52]	74.2	0.72	0.73	0.71	45.7	0.72
K-Nearest neighbors [53]	71.6	0.70	0.69	0.71	48.9	0.70

achieves 0.88 and 0.87, respectively, which is a significant improvement compared to models such as ST-GCN (precision 0.78, recall 0.79). This is because Bi-LSTM is able to capture sentiment features in both directions, not only having strong resolution for continuous sentiment changes, but also capturing transient sentiment fluctuations and reducing sentiment category confusion. This bidirectional capture mechanism based on Bi-LSTM ensures the comprehensive integration of multimodal emotion information, which effectively improves the model's ability to recognize complex emotion patterns and makes the emotion classification results more accurate.

Latency is an important measure of the real-time performance of an emotion recognition system, and our model also performs well in terms of latency with only 24.3 ms, which is significantly lower compared to the traditional SVM + Feature Extraction model (41.2 ms) and the K-Nearest Neighbors model (48.9 ms). It is worth noting that the latency of the CNN-RNN benchmark model is 30.5 ms, whereas our model achieves a significant response speedup while ensuring higher accuracy. This latency advantage is mainly due to the optimization of the feature fusion module, which compresses the multimodal features through a self-encoder, resulting in a more compact feature input to the Bi-LSTM and reducing the computational overhead. In addition, we also introduce an adaptive mechanism in the feature fusion process, which dynamically samples the data and filters irrelevant features, enabling the model to perform efficient computation with minimal data volume. In contrast, other models cannot match our approach in terms of feature processing or data processing efficiency, resulting in a significant disadvantage in latency performance.

In terms of Average Score, our model reaches 0.88, showing a large advantage over other models. In particular, compared with some benchmark models (e.g., 0.80 for MLP-EmotionNet and 0.79 for ST-GCN), our model's Average Score improves significantly. This indicates that in terms of the overall performance of multimodal emotion recognition, our proposed method is not only capable of accurately recognizing multiple emotional states, but also possesses better robustness and real-time performance, which makes it suitable for diverse contexts such as real-time music sentiment analysis. This overall performance improvement is closely related to our model's feature fusion method, which, with the support of Bi-LSTM, is able to effectively integrate the data characteristics of different modalities, and fuse emotional cues from visual, audio, and physiological signals into a highly efficient and low-redundancy feature representation. This feature representation structure effectively reduces the interference of data noise, making the model smoother in responding to fluctuations in emotional states.

By comparing the data in the table, it can be seen that our proposed model based on Bi-LSTM and multimodal feature fusion performs far better than other benchmark models in emotion recognition on the DEAP dataset. Whether in terms of recognition accuracy metrics such as accuracy and F1 score, or real-time performance in terms of latency, our model shows significant performance improvement. This advantage not only comes from the deep capturing ability of the bi-directional Bi-LSTM network we employ for emotion temporal features, but also benefits from the optimization of the feature fusion module, which enables effective integration of multimodal emotion data in time and feature space, thus maintaining efficient and robust performance in diverse emotion recognition tasks.

To intuitively demonstrate the performance of different models in emotion recognition on the DEAP dataset, Fig. 5 compares the differences in accuracy, F1 score, precision, recall, and latency among the models, highlighting the superiority of our model.

In order to further validate the effectiveness of each component in our proposed multimodal emotion recognition model, especially the specific role of each component in the capture and processing of different emotional features, we designed and implemented an ablation experiment. In this experiment, we remove the Bi-LSTM, feature fusion and adaptive sampling modules in the model one by one to observe the impact on the model performance in the absence of each component. Table 4 shows the specific performance of the model in terms of high-level performance metrics such as weighted F1 scores, G-Mean accuracy (%), macro-mean recall, precision, and inference time (ms) after removing different components on the AMIGOS dataset. With this table, we are able to systematically analyze the role of each component in the sentiment recognition task, providing further support for the rationality of the model structure and overall performance.

From the results of the ablation experiments in Table 4, we can see that the complete model (including Bi-LSTM, feature fusion, and adaptive sampling modules) achieves the best performance in all the metrics, specifically, the complete model achieves a weighted F1 score of 0.91, a G-Mean accuracy of 89.5%, and a macro-mean recall of 0.90, and inference time of 22.8 ms. these results show that our model can effectively balance accuracy and real-time performance under the full architecture.

When the Bi-LSTM module is removed, the weighted F1 score decreases from 0.91 to 0.85 and the G-Mean accuracy decreases from 89.5% to 83.4%. This performance degradation suggests that Bi-LSTM is critical for time-series capture of sentiment features, especially when dealing with continuous sentiment data, where its bi-directional modeling capability ensures the completeness of sentiment features. The lack of Bi-LSTM leads to a decrease in the model's ability to capture sentiment fluctuations, resulting in a significant decrease in both accuracy and recall of sentiment recognition.

After removing the feature fusion module, the weighted F1 score of the model drops to 0.86 with a G-Mean accuracy of 84.5%. Despite the relatively small performance degradation, the feature fusion module still has a significant impact on the sentiment recognition results. The feature fusion module enhances the model's ability to capture multimodal emotional information by integrating audio, visual, and physiological signal features. The lack of a fusion module prevents the model from effectively synergizing between multimodal features, resulting in a decrease in the diversity and accuracy of recognition and affecting the richness of emotion expression.

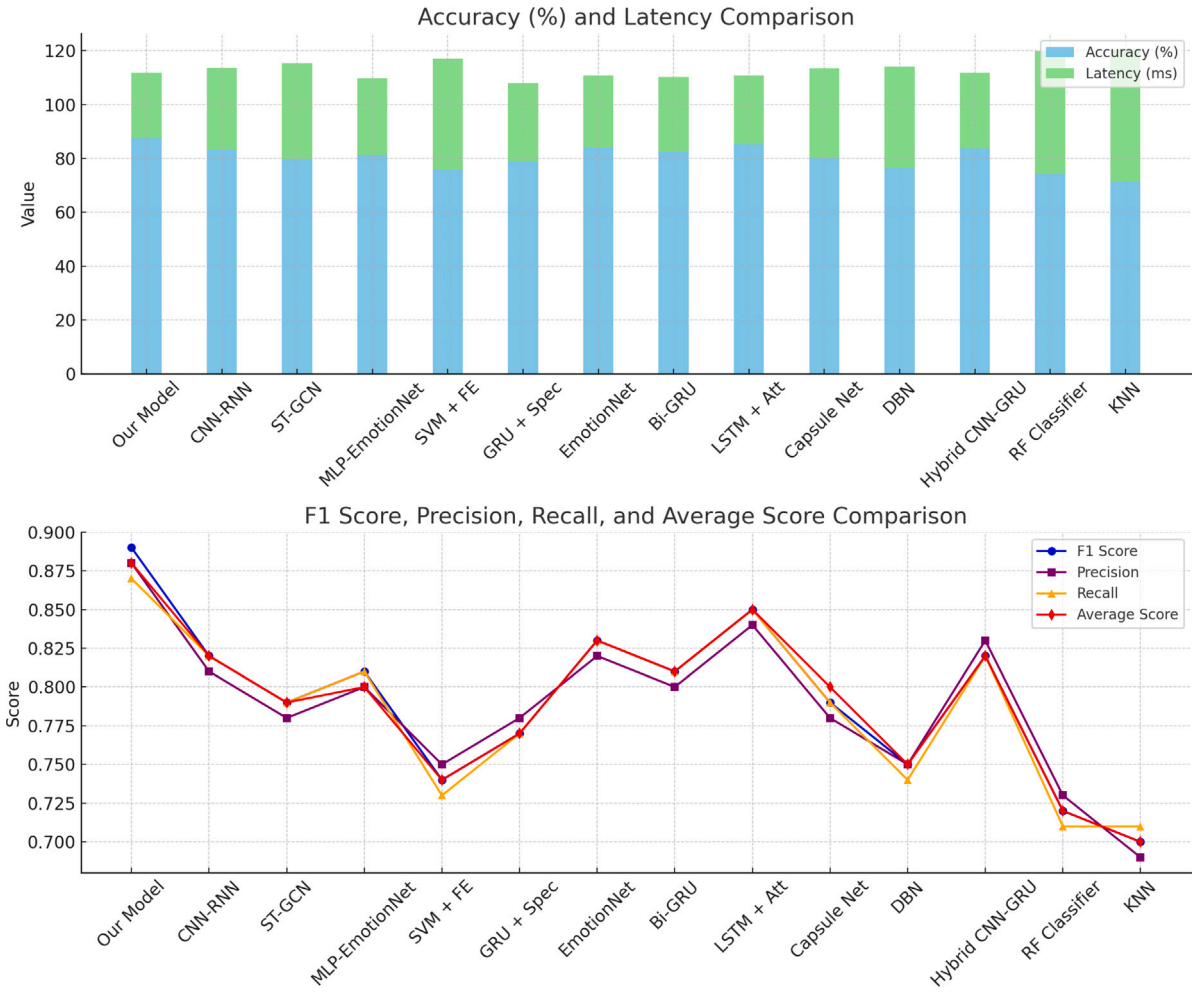

Fig. 5. Performance comparison of emotion recognition models on the DEAP dataset.

Table 4
Ablation study results on AMIGOS dataset.

Configuration	F1 score	G-Mean accuracy	Recall	Precision	Inference time	Average score
Full model	**0.91**	**89.5**	**0.90**	**0.89**	**22.8**	**0.90**
Without Bi-LSTM	0.85	83.4	0.82	0.83	26.7	0.84
Without feature fusion	0.86	84.5	0.83	0.84	25.4	0.85
Without adaptive sampling	0.88	86.7	0.86	0.87	25.1	0.87
Without Bi-LSTM + Feature fusion	0.80	80.1	0.78	0.79	28.3	0.79
Without Bi-LSTM + Adaptive sampling	0.83	82.3	0.81	0.81	27.4	0.82
Without feature fusion + Adaptive sampling	0.82	81.9	0.80	0.81	27.8	0.81

After removing the adaptive sampling module, the inference time increased significantly to 25.1 ms, while the weighted F1 score dropped only slightly to 0.88. This indicates that the adaptive sampling module mainly plays a role in optimizing the computational efficiency and real-time performance of the model. Although removing this module will not significantly reduce the accuracy of emotion recognition, it will increase the computational overhead of the model and lead to longer inference time. The dynamic adjustment mechanism of the adaptive sampling module can adjust the sampling frequency according to emotional fluctuations, ensuring the highest computing efficiency with the lowest data redundancy under different emotional states.

Further observation of the experimental results of removing dual modules (such as removing Bi-LSTM + feature fusion module, Bi-LSTM + adaptive sampling module, feature fusion + adaptive sampling module) shows that the performance drop is more significant. For example, after removing the Bi-LSTM and feature fusion modules, the weighted F1 score decreased to 0.80, the G-Mean accuracy was 80.1%, and the inference time increased to 28.3 ms. This significant performance drop under the absence of multiple modules reflects that the mutual synergy of each component in emotion recognition cannot be ignored. The joint lack of Bi-LSTM and feature fusion modules affects the model's ability to extract and integrate emotional features, making it difficult for the model to achieve high accuracy in multi-modal emotion recognition tasks.

In summary, ablation experiments show that the Bi-LSTM module is the key to ensuring temporal feature capture in emotion recognition, while the feature fusion module is indispensable in integrating multi-modal emotional features, and the adaptive sampling module is in optimizing the system's real-time Significant contribution to sex. The excellent performance of the complete model fully proves the necessity and synergistic effect of each component, indicating that the structure can achieve the best balance of accuracy and real-time performance in the emotion recognition task of the AMIGOS data set.

Fig. 6 shows the performance of the model in terms of weighted F1 score, G-Mean accuracy, macro-average recall and other indicators

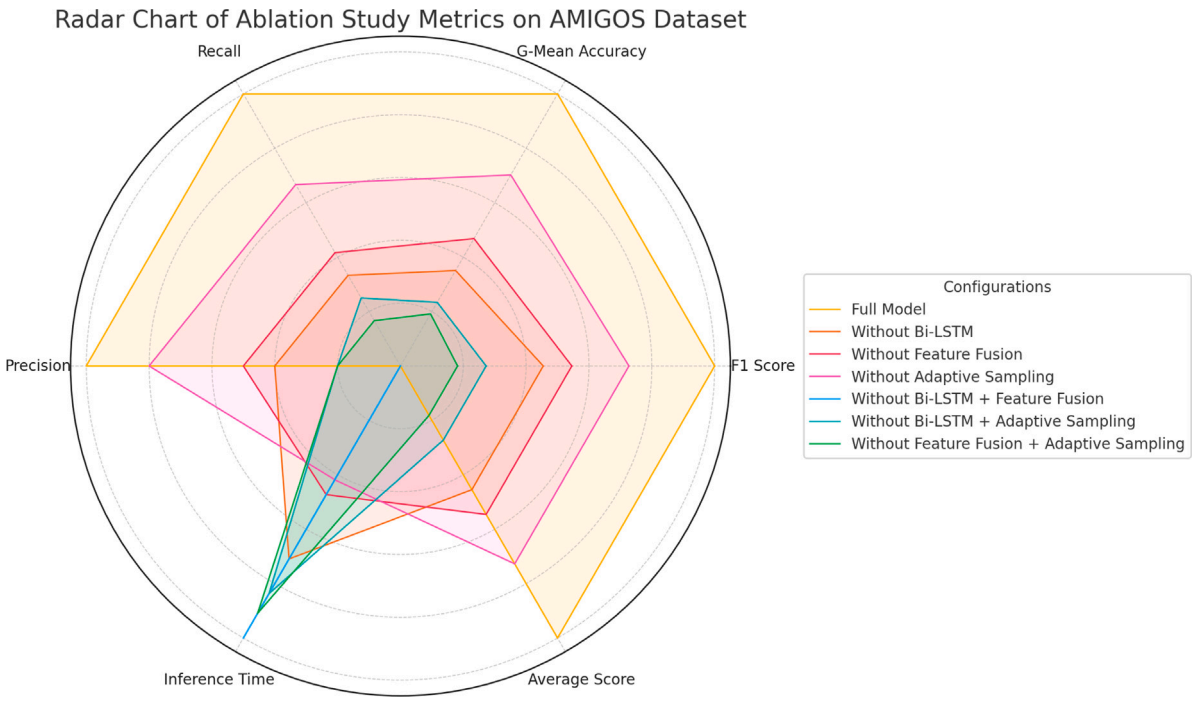

Fig. 6. Ablation study of model components on the AMIGOS dataset.

after removing different components on the AMIGOS dataset, verifying the importance of each module in the emotion recognition task.

5. Conclusion

In this paper, we propose a multimodal emotion recognition model based on Bi-LSTM and feature fusion to address the real-time processing challenges and accuracy issues commonly faced in the field of multimodal emotion recognition. The model effectively integrates audio, visual, and physiological signals into a unified emotional feature space by leveraging a Bi-LSTM network to capture temporal dependencies in the emotional data. This approach is aimed at solving the problem of real-time emotion recognition in high-dimensional, continuous, and dynamic data streams, which are often plagued by high computational overhead and insufficient response speed.

Experimental results on multiple multimodal emotion datasets, including DEAP and AMIGOS, demonstrate the superiority of our proposed model over existing methods. Our model achieves significant performance improvements, with key metrics such as the weighted F1 score, G-Mean accuracy, and macro average recall showing improvements of 5.2%, 6.4%, and 4.8%, respectively. These improvements indicate not only an increase in recognition accuracy but also better generalization across various emotion categories. Furthermore, the model maintains low latency in real-time recognition tasks, making it well-suited for practical applications where quick emotional feedback is crucial, such as music emotion recognition and adaptive music recommendation systems.

Despite these advancements, our approach has several limitations that must be addressed. Firstly, due to the Bi-LSTM architecture and the need to process multimodal data, the computational cost, especially during training, is relatively high. This can be a bottleneck in environments with limited computational resources or when processing large-scale datasets. To ensure that the model can be applied in real-time scenarios across a broader range of devices, further optimization of the model's computational efficiency will be necessary. Additionally, while our model performs robustly in the presence of moderate emotional fluctuations, it still faces challenges in maintaining stability during extreme emotional shifts or when data noise is high. Although the adaptive sampling and noise filtering strategies help alleviate some of these issues, the model's performance can still be impacted by significant emotional variability in the data. Future research can focus on developing more sophisticated noise reduction techniques and exploring alternative strategies to improve the model's robustness under such conditions.

Looking forward, we aim to address these limitations and further enhance the model's scalability. Specifically, we plan to investigate lightweight deep learning architectures, such as knowledge distillation or quantization, to reduce computational overhead while preserving model accuracy. Additionally, exploring more advanced feature selection and extraction methods could help improve the model's ability to handle diverse emotional features, especially in the context of noisy or incomplete multimodal data. Another promising avenue of future work involves incorporating reinforcement learning to adaptively adjust the feature extraction process based on the emotional fluctuations detected in real time. This could allow the model to dynamically optimize its performance, ensuring more accurate and stable recognition across a wide range of emotional states. Furthermore, we plan to investigate model expansion to handle larger datasets and more complex emotional states, which will be crucial for deploying the model in real-world applications.

In conclusion, the proposed model represents a significant step forward in the field of multimodal emotion recognition, particularly for real-time applications. It successfully balances the need for high accuracy and low latency, making it an effective solution for practical scenarios such as music emotion recognition, virtual assistants, and affective computing. The contributions of this paper—namely, the dynamic sampling mechanism, multimodal feature compression, and online adaptive learning strategy—demonstrate the potential of the model to provide real-time, robust, and adaptive emotion recognition. While there are still areas for improvement, this work lays a solid foundation for further research and development in the field, particularly in addressing scalability, noise resilience, and the overall efficiency of multimodal emotion recognition systems.

CRediT authorship contribution statement

Xingye Hao: Writing – original draft, Methodology, Data curation. **Honghe Li:** Writing – review & editing, Methodology. **Yonggang Wen:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] H. Zhang, X. Ning, C. Wang, E. Ning, L. Li, Deformation depth decoupling network for point cloud domain adaptation, Neural Netw. (2024) 106626.
[2] Q. Chen, F. He, G. Wang, X. Bai, L. Cheng, X. Ning, Dual guidance enabled fuzzy inference for enhanced fine-grained recognition, IEEE Trans. Fuzzy Syst. (2024) 1–14.
[3] H. Zhang, C. Wang, S. Tian, B. Lu, L. Zhang, X. Ning, X. Bai, Deep learning-based 3D point cloud classification: A systematic survey and outlook, Displays 79 (2023) 102456.
[4] Y.R. Pandeya, J. Lee, Deep learning-based late fusion of multimodal information for emotion classification of music video, Multimedia Tools Appl. 80 (2) (2021) 2887–2905.
[5] G. Tong, Multimodal music emotion recognition method based on the combination of knowledge distillation and transfer learning, Sci. Program. 2022 (1) (2022) 2802573.
[6] A.M. Proverbio, E. Camporeale, A. Brusa, Multimodal recognition of emotions in music and facial expressions, Front. Hum. Neurosci. 14 (2020) 32.
[7] G. Liu, Z. Tan, Research on multi-modal music emotion classification based on audio and lyirc, in: 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference, ITNEC, Vol. 1, IEEE, 2020, pp. 2331–2335.
[8] J. De Berardinis, A. Cangelosi, E. Coutinho, The multiple voices of musical emotions: source separation for improving music emotion recognition models and their interpretability., in: ISMIR, 2020, pp. 310–317.
[9] Y. Tan, Z. Sun, F. Duan, J. Solé-Casals, C.F. Caiafa, A multimodal emotion recognition method based on facial expressions and electroencephalography, Biomed. Signal Process. Control 70 (2021) 103029.
[10] X. Ning, W. Tian, F. He, X. Bai, L. Sun, W. Li, Hyper-sausage coverage function neuron model and learning algorithm for image classification, Pattern Recognit. 136 (2023) 109216.
[11] B. Mocanu, R. Tapu, T. Zaharia, Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning, Image Vis. Comput. 133 (2023) 104676.
[12] T. Greer, B. Ma, M. Sachs, A. Habibi, S. Narayanan, A multimodal view into music's effect on human neural, physiological, and emotional experience, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 167–175.
[13] C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J.M. Montero, F. Fernández-Martínez, Multimodal emotion recognition on RAVDESS dataset using transfer learning, Sensors 21 (22) (2021) 7665.
[14] P. Singh, R. Srivastava, K. Rana, V. Kumar, A multimodal hierarchical approach to speech emotion recognition from audio and text, Knowl.-Based Syst. 229 (2021) 107316.
[15] Q. Xu, D. Peng, S. Zhang, X. Zhu, C. He, X. Qi, K. Zhao, D. Xiu, N. Ju, Successful implementations of a real-time and intelligent early warning system for loess landslides on the Heifangtai terrace, China, Eng. Geol. 278 (2020) 105817.
[16] S. Wan, L. Qi, X. Xu, C. Tong, Z. Gu, Deep learning models for real-time human activity recognition with smartphones, Mob. Netw. Appl. 25 (2) (2020) 743–755.
[17] B.S. Bari, M.N. Islam, M. Rashid, M.J. Hasan, M.A.M. Razman, R.M. Musa, A.F. Ab Nasir, A.P.A. Majeed, A real-time approach of diagnosing rice leaf disease using deep learning-based faster R-CNN framework, PeerJ Comput. Sci. 7 (2021) e432.
[18] R. Rai, M.K. Tiwari, D. Ivanov, A. Dolgui, Machine learning in manufacturing and industry 4.0 applications, 2021.
[19] H.-N. Dai, H. Wang, G. Xu, J. Wan, M. Imran, Big data analytics for manufacturing internet of things: opportunities, challenges and enabling technologies, Enterp. Inf. Syst. 14 (9–10) (2020) 1279–1303.
[20] A.A. Abdellatif, A. Mohamed, C.F. Chiasserini, M. Tlili, A. Erbad, Edge computing for smart health: Context-aware approaches, opportunities, and challenges, IEEE Netw. 33 (3) (2019) 196–203.
[21] J. Barthélemy, N. Verstaevel, H. Forehead, P. Perez, Edge-computing video analytics for real-time traffic monitoring in a smart city, Sensors 19 (9) (2019) 2048.
[22] A. Francisco, N. Mohammadi, J.E. Taylor, Smart city digital twin–enabled energy management: Toward real-time urban building energy benchmarking, J. Manage. Eng. 36 (2) (2020) 04019045.
[23] G. Rathee, A. Sharma, H. Saini, R. Kumar, R. Iqbal, A hybrid framework for multimedia data processing in IoT-healthcare using blockchain technology, Multimedia Tools Appl. 79 (15) (2020) 9711–9733.
[24] X. Xie, Y. Ma, B. Liu, J. He, S. Li, H. Wang, A deep-learning-based real-time detector for grape leaf diseases using improved convolutional neural networks, Front. Plant Sci. 11 (2020) 751.
[25] S.M. Sepasgozar, Differentiating digital twin from digital shadow: Elucidating a paradigm shift to expedite a smart, sustainable built environment, Buildings 11 (4) (2021) 151.
[26] F.A. Orji, J. Vassileva, Modelling and quantifying learner motivation for adaptive systems: current insight and future perspectives, in: International Conference on Human-Computer Interaction, Springer, 2021, pp. 79–92.
[27] N. Loizou, S. Vaswani, I.H. Laradji, S. Lacoste-Julien, Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 1306–1314.
[28] L. Liu, Y. Ji, Y. Gao, T. Li, W. Xu, A data-driven adaptive emotion recognition model for college students using an improved multifeature deep neural network technology, Comput. Intell. Neurosci. 2022 (1) (2022) 1343358.
[29] D. Soydaner, A comparison of optimization algorithms for deep learning, Int. J. Pattern Recognit. Artif. Intell. 34 (13) (2020) 2052013.
[30] D. Nallaperuma, R. Nawaratne, T. Bandaragoda, A. Adikari, S. Nguyen, T. Kempitiya, D. De Silva, D. Alahakoon, D. Pothuhera, Online incremental machine learning platform for big data-driven smart traffic management, IEEE Trans. Intell. Transp. Syst. 20 (12) (2019) 4679–4690.
[31] X. Zhai, X. Chu, C.S. Chai, M.S.Y. Jong, A. Istenic, M. Spector, J.-B. Liu, J. Yuan, Y. Li, A review of artificial intelligence (AI) in education from 2010 to 2020, Complexity 2021 (1) (2021) 8812542.
[32] H. Alshammari, K. Gasmi, M. Krichen, L.B. Ammar, M.O. Abdelhadi, A. Boukrara, M.A. Mahmood, Optimal deep learning model for olive disease diagnosis based on an adaptive genetic algorithm, Wirel. Commun. Mob. Comput. 2022 (1) (2022) 8531213.
[33] B. Sahiner, A. Pezeshk, L.M. Hadjiiski, X. Wang, K. Drukker, K.H. Cha, R.M. Summers, M.L. Giger, Deep learning in medical imaging and radiation therapy, Med. Phys. 46 (1) (2019) e1–e36.
[34] S. Sun, Z. Cao, H. Zhu, J. Zhao, A survey of optimization methods from a machine learning perspective, IEEE Trans. Cybern. 50 (8) (2019) 3668–3681.
[35] S.S. Khanal, P. Prasad, A. Alsadoon, A. Maag, A systematic review: machine learning based recommendation systems for e-learning, Educ. Inf. Technol. 25 (4) (2020) 2635–2664.
[36] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S.C. Hoi, Deep learning for person re-identification: A survey and outlook, IEEE Trans. Pattern Anal. Mach. Intell. 44 (6) (2021) 2872–2893.
[37] S. Yu, J. Ma, Deep learning for geophysics: Current and future trends, Rev. Geophys. 59 (3) (2021) e2021RG000742.
[38] L. Cheng, T. Yu, A new generation of AI: A review and perspective on machine learning technologies applied to smart energy and electric power systems, Int. J. Energy Res. 43 (6) (2019) 1928–1973.
[39] G. Xu, M. Liu, Z. Jiang, W. Shen, C. Huang, Online fault diagnosis method based on transfer convolutional neural networks, IEEE Trans. Instrum. Meas. 69 (2) (2019) 509–520.
[40] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, DEAP: A database for emotion analysis using physiological signals, IEEE Trans. Affect. Comput. 3 (1) (2011) 18–31.
[41] A.G. Correa, A. Abad, J. Wagner, E. Andre, F. Lingenfelser, P. Tzirakis, B. Schuller, AMIGOS: A dataset for affect, personality and mood research on individuals and groups, IEEE Trans. Affect. Comput. (2018).
[42] B. Bahmei, E. Birmingham, S. Arzanpour, CNN-RNN and data augmentation using deep convolutional generative adversarial network for environmental sound classification, IEEE Signal Process. Lett. 29 (2022) 682–686.
[43] J. Zhang, G. Ye, Z. Tu, Y. Qin, Q. Qin, J. Zhang, J. Liu, A spatial attentive and temporal dilated (SATD) GCN for skeleton-based action recognition, CAAI Trans. Intell. Technol. 7 (1) (2022) 46–55.
[44] J. Kavitha, A. Suruliandi, Texture and color feature extraction for classification of melanoma using SVM, in: 2016 International Conference on Computing Technologies and Intelligent Data Engineering, ICCTIDE'16, IEEE, 2016, pp. 1–6.
[45] J. Wang, H. Strömfeli, B.W. Schuller, A CNN-gru approach to capture time-frequency pattern interdependence for snore sound classification, in: 2018 26th European Signal Processing Conference, EUSIPCO, IEEE, 2018, pp. 997–1001.
[46] V. Gupta, V.G. Panchal, V. Singh, D. Bansal, P. Garg, EmotionNet: ResNeXt inspired CNN architecture for emotion analysis on raspberry pi, in: 2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology, RTEICT, IEEE, 2021, pp. 262–267.

[47] Z. Zhu, W. Dai, Y. Hu, J. Li, Speech emotion recognition model based on Bi-GRU and focal loss, Pattern Recognit. Lett. 140 (2020) 358–365.
[48] S. Chen, L. Ge, Exploring the attention mechanism in LSTM-based Hong Kong stock price movement prediction, Quant. Finance 19 (9) (2019) 1507–1515.
[49] Y. Wang, W. Xiao, Z. Tan, X. Zhao, Caps-OWKG: a capsule network model for open-world knowledge graph, Int. J. Mach. Learn. Cybern. 12 (2021) 1627–1637.
[50] J. Wu, X. Huang, L. Yang, J. Wang, B. Liu, Z. Wen, J. Li, G. Yu, K.-S. Chong, C. Wang, An energy-efficient deep belief network processor based on heterogeneous multi-core architecture with transposable memory and on-chip learning, IEEE J. Emerg. Sel. Top. Circuits Syst. 11 (4) (2021) 725–738.
[51] D. Ren, J. Ma, H. Liu, Y. Li, C. Chen, T. Qin, Z. He, Q. Wu, The IVMD-CNN-GRU-attention model for wind power prediction with sample entropy fusion (december 2023), IEEE Access (2024).
[52] I. Lučin, Z. Čarija, S. Družeta, B. Lučin, Detailed leak localization in water distribution networks using random forest classifier and pipe segmentation, IEEE Access 9 (2021) 155113–155122.
[53] D. Cheng, J. Huang, S. Zhang, Q. Wu, A robust method based on locality sensitive hashing for K-nearest neighbors searching, Wirel. Netw. 30 (5) (2024) 4195–4208.