Multimedia Systems (2025) 31:278
https://doi.org/10.1007/s00530-025-01871-w

SURVEY

A survey on music emotion recognition using learning models

Yixin Wang[1] · Xujian Zhao[1] · Chuanpeng Deng[1] · Yao Xiao[1] · Haoxin Ruan[1] · Peiquan Jin[2] · Xuebo Cai[3]

Received: 13 July 2024 / Accepted: 24 May 2025 / Published online: 4 June 2025

Abstract
Music is one of the art forms for expressing emotions, and it conveys emotional information through elements such as combinations of notes, melodic variations, rhythmic modulation, and choice of timbre. Recently, music emotion recognition has attracted the attention of researchers due to that it can be widely applied under different scenarios, such as music recommendation systems, intelligent music generation and creation, music therapy and emotion regulation, and other fields. With the development of Artificial Intelligence, deep learning-based music emotion recognition technology has gradually replaced traditional machine learning technology, becoming the mainstream of the times. The purpose of this paper is to present a summary of current studies on music emotion recognition. Firstly, we introduce the task of music emotion recognition from some relevant definitions, processes, and emotion models. Then, the main current advances in music emotion recognition are detailed in terms of both traditional machine learning and deep learning. Next, some commonly used public datasets are presented, as well as evaluation metrics. Finally, we summarize the current research challenges facing music emotion recognition and the future trends.

Keywords Music emotion recognition · Music information retrieval · Machine learning · Artificial intelligence

1 Introduction

Since the beginning of the smart era, the trend of electronic music has been rapidly developing, allowing people to access a huge amount of music resources from the cloud. Classifying and managing these massive music resources enables people to quickly find the music works they need. As we know, music is a speechless language that can deeply touch people's emotions. It becomes especially important to accurately identify the emotional labels of a musical piece. Emotional labels can help listeners better understand and feel the music by associating it with a particular mood, atmosphere, or scene. However, manually assigning labels to music is a time-consuming and labor-intensive task. Meanwhile, considering the market scale and diversity of the electronic music market, relying on manual methods alone cannot satisfy the demand for fast and accurate labeling. Therefore, research on automatic identification of music emotion labels by utilizing artificial intelligence techniques has naturally emerged.

Communicated by Junyu Gao.

✉ Xujian Zhao
jasonzhaoxj@swust.edu.cn

Yixin Wang
yixin1202@mails.swust.edu.cn

Chuanpeng Deng
chuanpengdeng@mails.swust.edu.cn

Yao Xiao
xiaobaiyc@outlook.com

Haoxin Ruan
rhx1999@gmail.com

Peiquan Jin
jpq@ustc.edu.cn

Xuebo Cai
53979676@qq.com

[1] School of Computer Science and Technology, Southwest University of Science and Technology, 59 Qinglong Road, Mianyang 621010, Sichuan, China

[2] School of Computer Science and Technology, University of Science and Technology of China, 96 JinZhai Road, Hefei 230026, Anhui, China

[3] School of Music and Dance, Sichuan University of Culture and Arts, 83 East Road, Mianyang 621000, Sichuan, China

MER (Music emotion recognition) has been a major research subject in MIR (Music Information Retrieval) [1]. The purpose of MER is to identify the emotions expressed from music by extracting and analyzing musical features through the computer and forming mapping relationships between musical features and emotional space [2]. Different from general emotion recognition tasks that directly reflect a specific emotion in words or facial expressions, the emotion expression of music emotion recognition tasks is often more abstract and multi-dimensional. A song may contain multiple emotional elements simultaneously, and these emotions may change as the melody develops. Moreover, with the development of technology, the source of music features is not only limited to audio signals, but also includes music symbolic, lyrics text, and even physiological information of living beings. Traditional feature engineering and machine learning have limited performance in extracting complex features from multivariate time series information. Consequently, deep learning techniques are more popular due to their ability to provide high-level abstract features.

MER also has a wide range of applications, such as music psychotherapy, music recommendation and user taste analysis on iTunes, personalized instruction in music education, and more. MER's research also provides technical and theoretical support for the improvement of multimedia systems, giving multimedia systems the ability to understand users' emotional needs and capture users' preferences. MER is not only an important bridge connecting user emotion and multimedia content, but also one of the key technologies to improve the intelligence level of multimedia systems. It helps to create a richer and personalized user experience, and provides users with more in-depth content services with stronger emotional resonance, enriching the functions of the multimedia system and user experience. Therefore, the research of MER plays an essential role in multimedia systems.

The same lyrics may show different emotions in the case of different accompaniments. This is one of the reasons why research in the field of MER is going to gradually transfer from unimodal to multimodal. Emotional information from multiple modalities provides more comprehensive and rich feature information. Unimodal models, on the other hand, are difficult to provide complete and complex emotional information. The creation of emotional features involves various factors resulting from non-linear interactions between different modalities, and multi-dimensional data.

MER research dates back to 1930, when researchers began exploring the relationship between music and emotion. In 2007, the Music Information Retrieval Evaluation Exchange (MIREX), a closely watched international competition for audio retrieval and evaluation, introduced the Audio Sentiment Classification (AMC) as part of the competition. Although there has been a review of the research on MER, with the proposal of multimodal technology and large language model, the field of MER has ushered in new opportunities and challenges. These new technologies have the potential not only to revolutionize the way music emotions are understood, but also to greatly enrich the application scenarios of MER. Based on this, we hope to conduct this survey to summarize the current development status of MER and provide guidance for its future development direction.

The contributions of this paper are:

(1) This paper provides a systematic summary of research tasks, recent advances, challenges, and future research directions in the field of MER.
(2) For the core steps of the MER task: feature extraction and emotion recognition. Firstly, two mainstream approaches (traditional machine learning and deep learning) that are commonly used are introduced. Then, the research work in the field of MER based on the two approaches is summarized separately. This allows researchers to have a clearer understanding of the latest research work in the direction of music emotion recognition.
(3) Authoritative and publicly available datasets in the field of MER are introduced in detail. The evaluation metrics of classification and regression models for the performance measurement of MER are given separately, such as accuracy, recall, F1-score, R^2, and Root Mean Square Error.

The remainder of this paper is structured as follows: In Sect. 2, we provide an overview of the preliminary knowledge pertinent to MER. In Sect. 3, we discuss several commonly employed emotional classification models. In Sect. 4, we offer a comprehensive summary of existing work in the MER field. Then, section 5 reviews widely used datasets and evaluation methodologies. Section 6 explores various applications of MER. In Sect. 7, we address current challenges and future trends in MER research. Finally, Sect. 8 concludes the paper with a summary of key findings and conclusions.

2 Overviewof MER

2.1 Preliminary

Existing MER research can be broadly categorized into two types, one of which investigates the overall emotion of music, while the other investigates the process of change in music emotion.

Definition 1: Static music emotion recognition. It refers to assigning emotion labels to entire songs, which focuses

on the moods and emotional states conveyed by the song as a complete musical composition.

Definition 2: Dynamic music emotion recognition. It analyzes the emotion-changing process of a song over a time series. Furthermore, it focuses on the evolution and changing trends of emotions throughout the song.

2.2 Workflow of MER

The MER task consists of the following three main components: data annotation, feature extraction, and emotion recognition. The basic flow is shown in Fig. 1.

(1) Data annotation: Based on research in music theory and emotion modeling, an appropriate emotion model is selected. Then, after listening to a short piece of music, human subjects annotate the emotions elicited by the music according to the selected emotion model.
(2) Feature extraction: Extraction of emotionally relevant and useful features through traditional machine learning or deep learning methodologies.
(3) Emotion recognition: Emotion labels are predicted by matching extracted features with subjective annotations provided by humans to evaluate model performance.

3 Data annotation

Data is fundamental to model training, especially for the complex music emotion recognition task. High-quality, diverse data sets can greatly affect the performance and generalization ability of models. Sufficient data volume, accurate and consistent labeling, and coverage of a wide range of scenarios ensure model performance. Therefore, precisely labeled data sets are an indispensable step in MER research. Emotion is an internal subjective experience, a psychological response that humans have to external stimuli or internal feelings. It is a state of mind and physical state produced by combining multiple feelings, thoughts, and behaviors [3]. In MER research, rationally annotating emotions is a complex and challenging problem. Due to the complexity and diversity of emotions, scholars have also proposed different emotion models as a basis for data annotation.

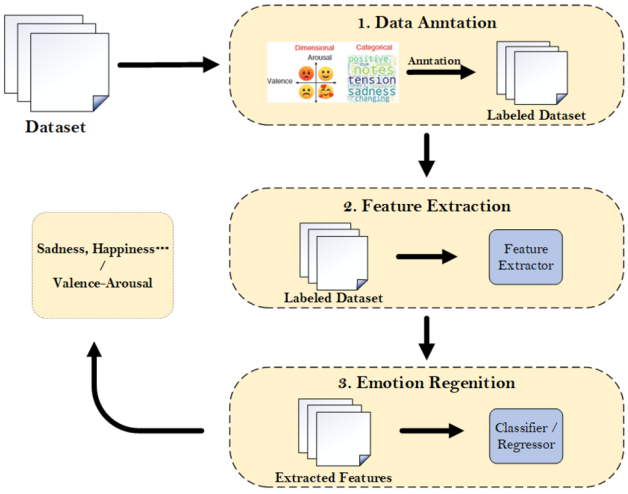

Fig. 1 MER framework

In Table 1, we summarize the emotion models commonly used in the MER task. Emotion models are typically classified into two groups: "General" and "Music". The "General" category encompasses emotion models that can be applied to emotion analysis tasks in various domains. On the other hand, the "Music" category comprises specifically designed music emotion models that more authentically capture the emotions when experiencing music.

Furthermore, in academia, two different models of emotion exist: "categorization" and "dimensionality", as shown in Fig. 2. Dimensional emotion models view emotions as continuous values on several dimensions, usually including positivity, negativity, arousal, and support. Conversely, categorical emotion models categorize emotions into predefined categories, such as "like," "dislike," "happy," "sad," and so on. Some scholars believe that the categorical emotion model is ambiguous, so dimensional emotion models have been used more recently [4–6].

Among the emotion models listed in Table 1, Hevner's affective ring [7] is one of the earliest and most influential emotion models of music. Hevner identified 67 affective adjectives to describe the emotional space of musical expression. These 67 emotional adjectives can be categorized into eight categories, namely dignified, sad, dreamy, serene, elegant, happy, euphoric, and dramatic, as shown in Fig. 3a. In addition, Russell's circumplex model of affect [8] is currently the most widely used emotion model in MER tasks. The model includes two dimensions: valence, which signifies the extent of positive and negative emotions, and arousal, which represents the intensity of emotion, the details are shown in Fig. 3b. Similarly, the emotion model created by Thayer [6], as shown in Fig. 3c, is a two-dimensional one, comprising energetic arousal and tense arousal. According to Thayer, valence is explicable by the varied amalgamations of energetic and tense arousal. Tellegen et al. [9] revised and supplemented the Thayer emotion model and proposed the TWC model, which used 38 adjectives to describe different emotions. It added a new coordinate system based on the original two-dimensional coordinate system, which has the characteristics of the natural and smooth emotional transition of the Thayer emotion model and greatly enriches the description of emotions, as shown in Fig. 3d. Furthermore, PAD [10] is a three-dimensional emotion model first proposed by Mehrabian and Russell detailed in Fig. 3(e). The

Fig. 2 Emotion models

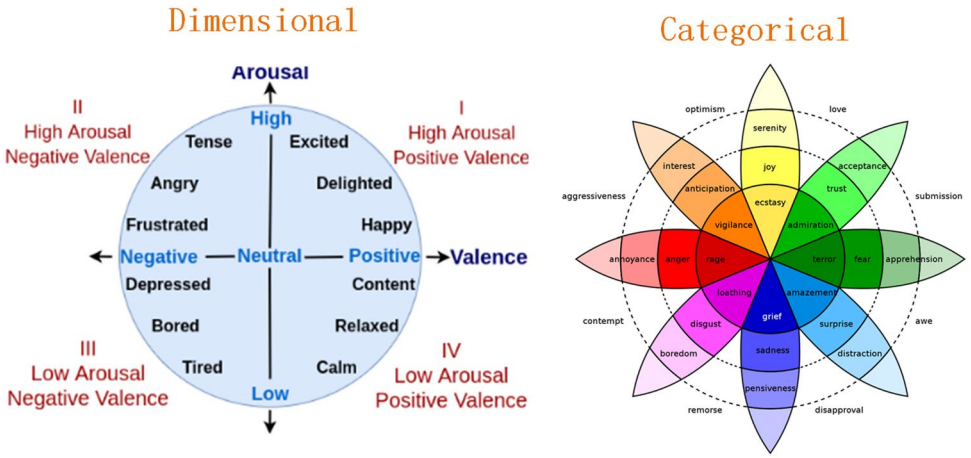

Table 1 Emotion models

Model name	Application domain	Emotion conceptualization	Classes/dimensions
Hevner affective ring [7]	Music	Categorical	67 classes
Russell's circumplex model [8]	General	Dimensional	2 dimensions
Thayer [6]	General	Dimensional	2 dimensions
TWC [9]	General	Dimensional	2 dimensions
PAD [10]	General	Dimensional	3 dimensions

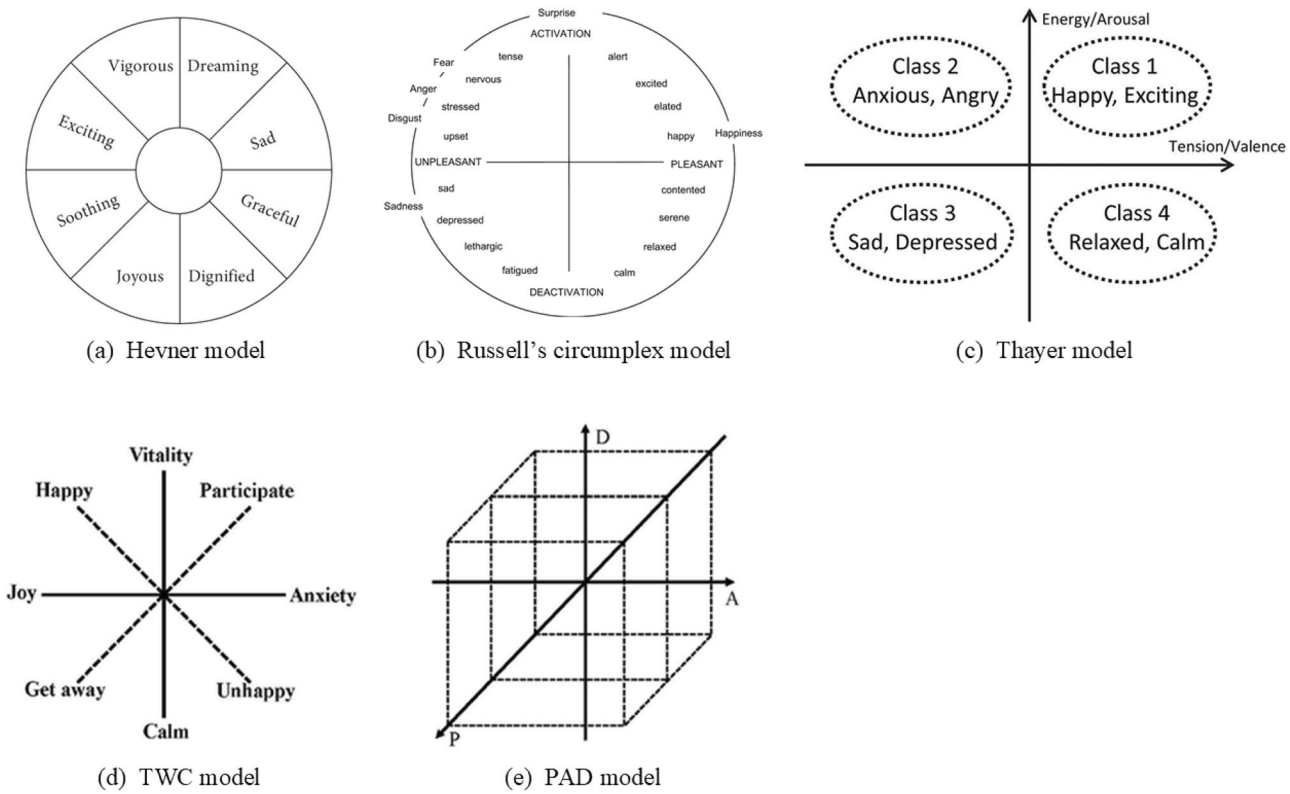

Fig. 3 Schematic diagram of the emotion models

model divides emotions into three dimensions: Pleasure (P), Arousal (A), and Dominance (D). Pleasure represents the positive and negative characteristics of an individual's emotional state, Arousal denotes the level of individual neurophysiological activation, and Dominance is the individual's control over the situation and others.

In addition to the emotion models described above, there are also some more specific ways of describing musical emotions. Such as probability distributions, antonym pairs, etc. In the following section, we will summarize the existing MER work in terms of both machine learning methods and deep learning methods.

4 Feature extraction and emotion recognition

Both traditional Machine Learning (ML) and Deep Learning (DL) methods have been used for MER but with slightly different approaches and results. ML-based methods extract hand-crafted features from recorded audio and use data-driven methods to predict the emotion of the music [11–13]. While DL-based methods can automatically extract features and complete the prediction process.

4.1 Traditional machine learning methods

The traditional MER method of machine learning involves two stages as shown in Fig. 4: extraction of hand-crafted features and prediction of emotion. We discuss them independently in Sects. 4.1.1 and 4.1.2, respectively.

4.1.1 Extraction of hand-crafted features

For the MER task, feature extraction is a fundamental problem. The quality of the extracted features will directly impact the model's accuracy in emotion recognition. The sources of features are diverse, such as audio signals, symbolic scores, lyrics, and even physiological data generated by the listener can be sources of features.

Audio feature: Audio features are the earliest and most commonly used features in the field of MER. In the past period, researchers depended on existing toolkits to extract audio features from waveform files [14, 15]. Wen et al. divided the emotions-related audio features into rhythmic, timbral, and spectral features [16]. Here rhythm encompasses musical features like pitch, duration, and intensity, which determine whether the sound is mellow and natural or not. Moreover, the timbre is a description of sound quality, and according to Chen et al. 's work [17] using timbre as an individual feature achieves better results in MER systems. Additionally, spectral features indicate the correlation between changes in the shape of the vocal tract and articulator emotion. Therefore, variations in the emotional content of the audio can greatly affect the spectral energy distribution in each spectral interval [18].

Lyric feature: Lyrics play a crucial role in communicating semantic information. In light of the rapid advancements in natural language processing technology, several studies have investigated feature extraction at the lyric level. Hu et al. conducted a comparative analysis of the effectiveness of BOW (bag of Words), POS (part-of-speech), and function words when extracting features from lyrics [19]. In addition, Hu et al. [20] proposed an emotion detection method for Chinese lyrics based on emotion vocabulary. Dakshina et al. [21] introduced an emotion recognition system for songs based on the Latent Dirichlet Allocation (LDA) modeling technique.

Symbolic feature: Symbolic features refer to the features extracted from the musical notation. In the existing research, there are fewer studies on this aspect. The MIDI (Musical Instrument Digital Interface) file is a music file format that directly contains precise sequences of pitches,

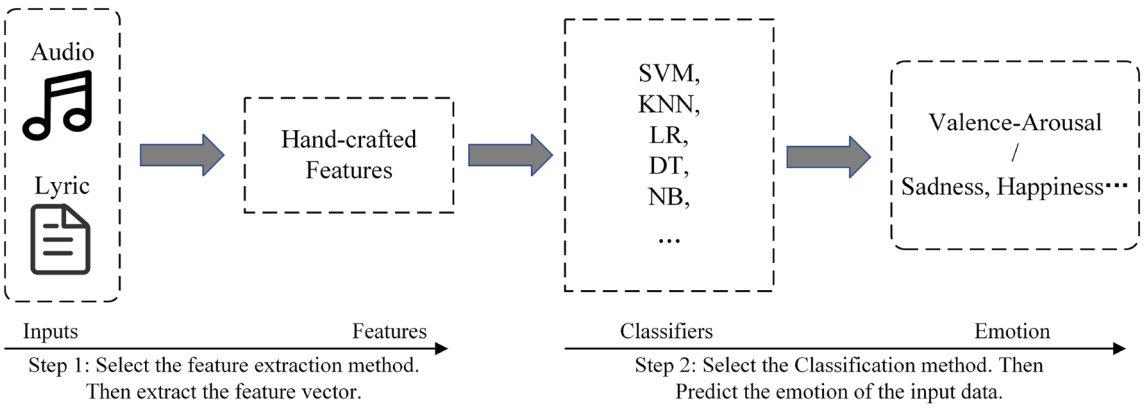

Fig. 4 The basic framework of traditional machine learning

intensities, etc., and is often employed to represent symbolic music scores.

Biological feature: Other than the audio, lyrics, and symbolic music scores mentioned above, physiological data collected from the listener after listening to the music can also serve as features for emotional recognition. Thammasan et al. [22] proposed a study to improve emotion recognition in music listening based on deep belief network (DBN) by utilizing electroencephalography (EEG) to effectively capture the characteristics of brain emotional information. With the development of medical technology and the emergence of wearable devices, the acquisition of physiological signals has become convenient, and the extraction of physiological features has been facilitated [23]. In addition, some researchers use unconventional methods of data collection. For instance, Nawa et al. [24] utilized functional magnetic resonance imaging (fMRI) to investigate the recognition of music emotions.

4.1.2 Prediction of emotion based on ML

After obtaining the hand-crafted features of the music, the next step is to select a suitable traditional machine-learning model for training and classification. The outcomes will vary based on the selected model. The methods commonly utilized in machine learning in the field of MER can be divided into two categories depending on the specific task: regression methods and classification methods. Table 2 shows some classical machine learning methods and Table 3 summarizes some representative works of the machine learning model.

Methods based on regression techniques are widely used in MER research. Specifically, Wang et al. [11] proposed a new Acoustic Emotion Gaussian (AEG) model to recognize emotions in music, which defines a proper generative process of emotion perception in music and demonstrates superior performance than the Support Vector Machine (SVM) and Multiple Linear Regression (MLR) models. Based on

Table 2 Several classical machine learning methods applied in MER

Model	Characteristics	References
LR	Simple structure, easy to understand, effective in small sample data sets, poorly fit for nonlinear relationships, and sensitive to outliers	[25] et al
SVM	Excellent performance in classification tasks, long training time, good generalization, difficult to optimize parameters, sensitive to data distribution	[23, 26] et al
KNN	Non-parametric model, easy to understand and implement, high computational cost, sensitive to noise, difficulty in choosing K value	[23] et al
DT	Simple structure, easy to understand, nonlinear decision making, missing value processing, easy to overfit, and lack global optimization	[23] et al
GPR	Suitable for small sample data, uncertainty quantification, fitting nonlinear relationships, complex hyperparameter tuning, and sensitive to kernel function selection	[27, 28] et al
NB	Simple and easy to understand, suitable for multi-classification tasks, effective for small samples, and sensitive to rare data	[23] et al

Table 3 MER methods based on traditional machine learning

Reference	Year	Feature modalities	Learning model	Emotion model	Dataset
[11]	2012	Audio	AEG	VA model	MTurk, MER60
[30]	2013	Audio	GPR	VA model	MediaEval Emotion
[25]	2014	Audio	LR	VA model	AMG240
[27]	2014	Audio	SVR, MLR, GPR	VA model	Self-built
[12]	2014	Audio, Lyric	NB	122 classes	Self-built
[26]	2014	Audio	SVM	4 classes	Self-built
[32]	2015	Audio	GPR	9 classes	Self-built
[33]	2015	Audio	CLR	6/18classes	EMOTIONS, CAL500
[28]	2016	Audio	GPR	VA model	MediaEval Emotion
[13]	2016	Audio	DS-SVR	VA model	MediaEval Emotion
[29]	2017	Audio	AEG	VA model	AMG1608
[31]	2018	Lyric	SVM	VA model	Self-built
[23]	2018	Physiological signals	SVM, NB, KNN, DT	3 classes	Self-built

the AGE model, Cheng et al. [29] investigated how to apply the AEG model to a personalized MER model with minimum user burden. Markov et al. [30] achieved favorable outcomes through the utilization of Gaussian Process Regression (GPR) to model arousal and valence. Chen et al. [25] proposed an adaptive approach based on Linear Regression (LR) to explore personalized MER tasks and demonstrated the effectiveness of the approach through comprehensive experiments.

Static and dynamic MER methods also have a lot of comparative exploration work. Soleymani et al. [27] tested static and dynamic methods for emotion recognition separately and discovered that Gaussian Process Regression (GPR) and Support Vector Regression (SVR) were more effective than MLR for static tasks. Xianyu et al. [13] proposed a new dual-scale support vector regression (DS-SVR) method for dynamic music emotion detection. The method set up two independent SVR models and decoupled two scales of emotion dynamics apart, one recognizes emotion changes between different songs and the other detects emotion changes within one song, and then combines the results of the two SVRs into the final result.

In addition, some researchers have made improvements in the inputs to the model. Fukuyama et al. [28] considered the acoustic signal of music as input for emotion recognition and conducted experiments with Gaussian Process Regression (GPR). Furthermore, Malheiro et al. [31] selected lyrics features instead of audio features as inputs for their model. In contrast to previous studies, they introduced three novel lyric features: slang presence, structural analysis, and semantic features.

Meanwhile, classification approaches have also been widely explored in the MER field. Specifically, Liu et al. [33] proposed an algorithm known as Multi-Emotion Similarity Preserving Embedding (ME-SPE) and combined it with Calibrated Label Ranking (CLR) to recognize the emotions of music. Additionally, Chiang et al. [26] demonstrated a hybrid framework that includes a kernel-based class separability (KBCS) feature selection method, a nonparametric weighted feature extraction (NWFE) method, and a hierarchical support vector machine (SVM) classifier for recognizing four types of music emotions: happy, nervous, sad, and calm.

Some work focuses on the inputs to the model. Wu et al. [12] introduced a Hierarchical Music Emotion Recognition (HMER) model, namely a hierarchical Bayesian model using sentence-level music and lyrics features, which captures music emotion dynamics through a song-segment-sentence hierarchal structure. In addition, Chen et al. [32] dealt with the classification problem from a regression perspective. They presented a music emotion detection system based on the deep Gaussian process (GP). In the feature extraction part, rhythm, dynamics, timbre, pitch, and intonation, which are five emotion-related features, are chosen to represent the music signal. Furthermore, Hu et al. [23] gathered physiological signals from study participants to create a dataset, which served as inputs for four distinct classification methods, namely SVM, Naive Bayes (NB), K-nearest neighbors (KNN), and Decision Trees (DT). Among these methods, KNN exhibited superior performance, surpassing the other three methods.

Note that there are numerous works on MER research based on traditional machine learning, with differences in their research backgrounds, theoretical foundations, and algorithmic performances. These works can be broadly categorized into two types: regression-based methods and classification-based methods. Although researchers have explored these two methods and achieved considerable performance, there are still shortcomings in traditional machine learning methods, such as difficulty in feature engineering, limited model generalization ability, and inability to handle long-term time-series information. Therefore, researchers have gradually shifted their attention to the field of deep learning, attempting to use the powerful modeling capabilities of neural network models to make up for the shortcomings of machine learning models.

4.2 Deep learning methods

Deep learning has developed rapidly in recent years since it was studied by the academic community in 2006. The proposal of DL models, such as Convolutional Networks (CNN) and Recurrent Neural Networks (RNN), has revitalized the development of the MER field. In Table 4, we summarized some typical deep learning models. They can be adopted as an end-to-end processing framework, capable of achieving the entire process of mapping raw data to the expected output. Compared to traditional machine learning models, deep learning-based MER models offer several advantages:

(1) Automatic feature extraction. ML methods typically require manual feature selection and extraction. In contrast, DL models learn abstract features from data through multi-layer neural networks, reducing the need for manual feature engineering.
(2) Powerful representation capability. DL models usually have stronger representation capability to model music emotions more complexly and accurately. The multi-layer structure and large number of parameters of deep learning models can help the models learn richer music features, improving the performance of emotion recognition tasks.
(3) Ability to build time series models. Music is a time-series data, and DL models are naturally suited to deal with time series. By adopting structures such as RNN

Table 4 Several topical deep learning methods applied in MER

Model	Characteristics	References
CNN	Extract key spatio-temporal feature information, but fail in extracting temporal dynamic information	[34, 35] et al
RNN	Capturing temporal dynamic information, prone to gradient loss, and difficult to handle long-term dependencies	[6] et al
LSTM	Efficiently handle long-term dependencies in sequence data with multiple gating mechanisms and complex training processes	[34, 36] et al
VGG	Extract high-level image features, unified 3x3 convolution kernel, huge number of parameters	[37] et al
GRU	Simple structure, fast training speed, suitable for short-term and medium-term dependencies, and low computational cost	[38] et al
GCN	Capture the relationship and structural information between nodes, effectively handle sparse data and irregular structures, and have high requirements for graph feature engineering	[39] et al
Transformer	Captures relationships between sequences, handles long-range dependencies, and supports parallel training	[40] et al

or CNN, DL models can effectively model the temporal relationships in music and better capture the changes in music emotion over time.

In sections 4.2.1 and 4.2.2, we will describe the application of DL in MER from two aspects, unimodal and multimodal, respectively. The basic frameworks of unimodal deep learnig and multimodal deep learnig are shown in Fig. 5 and Fig. 6.

4.2.1 Unimodal deep learnig

CNNs are one of the most representative learning models in the field of DL, which imitate the visual system of living beings through convolutional, pooling operations to learn representational features from the input data. In the MER task, CNN-based methods are widely applied to model learning. Liu et al. [34] transformed the raw audio into a corresponding spectrogram by applying a short-time Fourier transform, which was then used as input for a convolutional neural network (CNN). Each music spectrogram underwent convolution, pooling, and hidden layers before being predicted by the softmax layer. By using the CNN approach, no additional effort in extracting specific features was required. Experiments on the CAL500 and CAL500exp datasets showed that the proposed method obtains the optimal performance. Keelawat et al. [35] fed electroencephalogram (EEG) and other biological features as inputs to CNNs for emotion recognition and obtained results superior to methods such as SVM. In contrast to previous EEG-based work, they focused on creating a model that is subject-independent.

Furthermore, Chowdhury et al. [37] introduced a VGG-style deep neural network that leverages intermediate-level perceptual features, which represent various musical qualities and are understandable by human listeners without music theory knowledge. The proposed VGG-based model learned two independent tasks: the first is to forecast

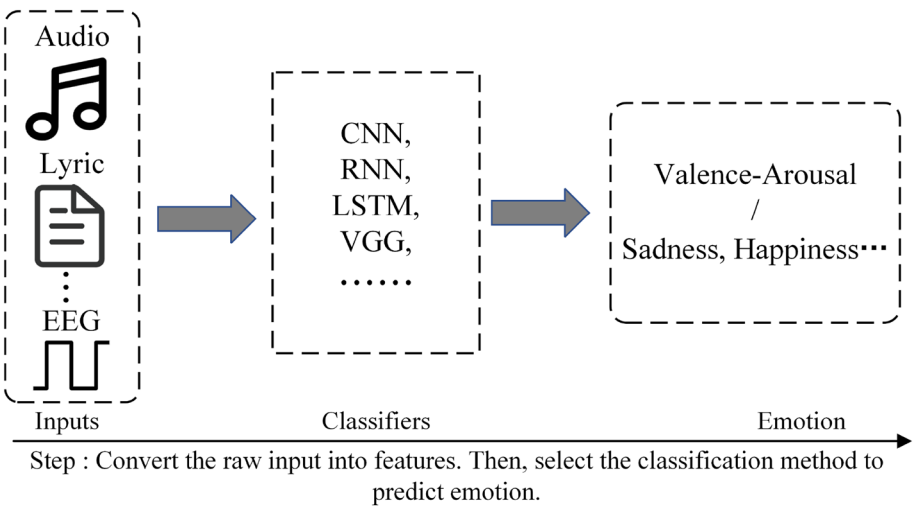

Fig. 5 The basic framework of unimodal deep learnig

mid-level features from the audio, while the second is to forecast emotions from the mid-level features. Yang et al. [41] presented a CNN-based model for recognizing emotions. Unlike other methods, they experimented with various feature extraction techniques. Among these techniques, the spectrogram based on the Constant-Q transform got the best performance. A deep neural network-based approach was proposed by Orjesek et al. [42], which mined music emotion-related salient features directly from the raw audio waveform. The method comprised stacked one-dimensional convolutional layers, an autoencoder layer based on iterative reconstruction, and bi-directional gated recursive units, resulting in outstanding performance in arousal and valence.

In addition, RNN has been widely used in MER because of its excellent sequential data processing capability. RNN has a feedback connection that allows information to be passed from the current time step to the next, which makes it memory-capable and able to utilize information from its history to influence the current output. Specifically, Weninger et al. [6] segmented the music into seconds and extracted features from each segmented segment, then fed the obtained features back into the LSTM to obtain the dynamically changing values of arousal and valence. The results of the experiments showed that LSTM outperforms SVR and feedforward neural networks in both continuous-time and static music emotion regression. Li et al. [43] believed that the emotion of music at a certain time scale is related to the temporal context and hierarchy of that scale. Therefore they proposed a multiscale regression method based on Deep Bidirectional Long Short-Term Memory (DBLSTM) to obtain temporal context information and hierarchy information. Their model was composed of three parts: the DBLSTM, the post-processing module, and the fusion module. The DBLSTM extracted temporal context information and hierarchical structure information from two directions and made arousal and valence predictions. The post-processing component further enhanced the temporal context processing capability for individual DBSLTM outputs. The fusion component integrated all DBLSTM models' outputs at different scales to obtain a single prediction. It had been experimentally demonstrated that DBLSTM was capable of capturing contextual information and achieving superior results compared to SVR and MLR. Moreover, Liu et al. [36] adopted BiLSTM (Bi-directional Long Short Term Memory) to extract features from audio. In addition to that, they also introduced a labeling space defined by antonym pairs, which made emotion labeling relatively objective. Kumar et al. [44] proposed using an Enhanced Residual Gated Recurrent Unit (RGRU) to extract data features, and the Adaptive Red Fox Algorithm (ARFA) to optimize RGRU hyperparameters to improve the accuracy of emotion recognition. And verify its performance on the EMOPIA dataset. Chang et al. [45] explored RNN networks of different complexity on different datasets. The experimental results showed that for small to medium-sized data sets, simpler models can provide good generalization capabilities, while for large-scale data sets, more complex or deep models are required to capture information. At the same time, the study also revealed the potential of neural networks in creating more personalized and emotionally resonant music recommendation and treatment systems.

Additionally, some researchers consider combining the feature extraction capability of CNNS with the sequential processing capability of RNNS to study MER. For example, the model called bidirectional convolutional recurrent sparse network (BCRSN) proposed by Dong et al. [46] was based on convolutional neural networks and recurrent neural networks. The model combined the advantages of CNN to learn features adaptively and LSTM to process sequence data more advantageously. The ability to learn features from spectrograms is enhanced by combining CNN and LSTM. An emotion prediction method that employs Convolutional Long Short-Term Memory Deep Neural Network (CLDNN) presented by Hizlisoy et al. [47]. In addition to utilizing conventional acoustic features, they incorporated log mel filter bank energies and mel frequency cepstrum coefficients (MFCCs), which were derived from audio data, as inputs to the network. Experiments showed that adding new features to standard acoustic features improves the model's classification performance. To capture the intricate emotional nuances in music, Yakovyna et al. [48] introduced a novel architecture, namely, hybrid CNN-LSTM. The architecture leverages CNN for robust feature detection and LSTM for effective sequence learning to address the temporal dynamics of music features, and achieves remarkable results in Emotify through meticulous fine-tuning strategies.

Some researchers have focused on the combination of attention mechanisms with RNN. The attentional LSTM model, proposed by Chaki et al. [49], combines a modified attentional mechanism with an LSTM. In this study, the emotion of music at a moment was determined by the music that came before that moment. Therefore, they only considered hidden information up to that moment. Multi-scale Context-based Attention modeling (MCA) is proposed by Ma et al. [50]. They used attention mechanisms to dynamically integrate temporal and hierarchical information from different time scales. In the proposed model, two attention mechanisms were designed. First, they trained an LSTM for each time scale to compute the arousal and valence value, and the first attentional mechanism would be used to compute the weights for each time step, and all the hidden layers would be weighted and summed to yield the final result for that time scale. Then, the second attention mechanism was used to calculate the weight values for each time scale, and then all the time scales were weighted and summed to obtain the final prediction. Integrating various time scales

employing attention mechanisms can facilitate acquiring dynamic music structure representations. In addition, to explore the emotional expressiveness of the violin, Ma et al. constructed a dataset dedicated to emotion recognition in violin music called VioMusic, and also proposed a baseline model called CNN-BiGRU-Attention (CBA), which utilizes CNNs to capture deep emotional features, BiGRUs to decode contextual relationships between musical emotions, and attention mechanisms to focus on the most emotional elements in the music.

As the attention mechanism demonstrated its powerful ability to extract features and focus on key information, researchers introduced it into MER tasks to improve MER model information extraction capability. Such as Agrawal et al. [51] presented a transformer-based approach for music emotion recognition that took the lyrics as input. The XLNet transformer model yielded the raw hidden states, which were then processed by the SequenceSummary block. This block generated a sequence of individual vector summaries of the hidden states, which were ultimately predicted by a fully connected layer. While MT-SMNN designed by Qiu et al. [52] was a transformer-based multi-tasking framework that concentrates on symbol-based recognition of emotions in music. The framework integrated emotion recognition with key classification and velocity classification tasks and experimentally confirmed the effectiveness of both additional tasks. Experiments demonstrated that MT-SMNN achieves superior performance on both EMOTPIA and VGMIDI datasets. Zhang et al. [53] introduced a model called Modular Composite Attention Network (MCAN), which compensated for the shortcomings of previous MER methods that were unable to extract representative emotion-related features, due to mainly utilizing simple convolutional layers to extract features from the raw audio signals. In addition, they introduced the sample reconstruction technique and the style embedding module to enhance the stability and detail handling of the network, respectively.

In addition, some studies have introduced special methods to assist MER research. Genre-Affect Relationship Network(GARN), is a multi-task learning framework proposed by Cheng et al [54]. The main task of the framework was emotion regression, and genre classification was the auxiliary task. The model embedded the relationship between effect and genre as tensor normal prior within task-specific layers, further optimizing the architecture by adding uncertainty. The results showed that adding an auxiliary task to learn the relationship between emotion and genre could benefit the emotion recognition process. Zhang et al. [55] proposed the CL-RL model by combining Curriculum Learning (CL) and Reinforcement Learning (RL). By integrating the capabilities of CL into RL, the proposed CL-RL model exhibits the ability to understand human thinking, thus solving the difficulty and complexity of understanding emotional feelings in the MER task. Chang et al. [56] introduced a new method for music emotion recognition. Specifically, the method uses intra-feature and inter-feature orthogonal fusion (IIOF) techniques to integrate local and global features learned from the original waveform of the music signal. By using a two-dimensional representation with different receptive fields, the model can capture multi-level embedding of time-frequency information. Subsequently, IIOF is used to enhance the complementarity between these features, reduce redundancy, and improve feature diversity. Han et al. [57] integrated physiological data obtained from wearable devices, including heart rate variability (HRV) and galvanic skin response (GSR), with music emotion classification to propose a three-dimensional emotion model. This model refines emotion categories and addresses the inconsistencies between traditional two-dimensional emotion models and psychological research findings. Additionally, a high-precision three-dimensional emotion classification model was designed based on the GAI architecture, which integrates convolutional neural networks (CNNs) and variational autoencoders (VAEs). The model incorporates multi-scale parallel convolutions and attention mechanisms to enhance the understanding of audio features.

In general, MER based on unimodal deep learning has been explored on CNNs, RNNs, and their variants, and all of them have achieved outstanding performance. In addition, the success of attention mechanisms has brought new research directions to the field of MER. Table 5 is a summary of relevant unimodal DL approaches in the MER field. To be sure, unimodal based MER work has yielded excellent results, but the single data source limits the upper limit of the model. Therefore, multimodal techniques have been introduced into MER tasks.

4.2.2 Multimodal deep learning

In recent years, with the development of multimodal technology, there are also some studies applying multimodal technology to the MER field. Compared to unimodal techniques, multimodal techniques that combine multiple types of data as inputs can provide richer semantic features, which can provide a more comprehensive and accurate understanding of musical emotions. Furthermore, inputs from different modalities can complement each other to provide more comprehensive emotional information for the model. Meanwhile, the multimodal approach is more robust to noise and interference.

Audio and lyrics are the two main components of music. Therefore, several researchers have conducted different studies on audio and lyrics. For example, the Bimodal Deep Boltzmann Machine (BDBM), proposed by Huang et al. [59], comprised a two-layer Deep Boltzmann Machine network: one for audio data and the other

Table 5 MER methods based on the unimodal deep learning

Reference	Year	Feature modalities	Learning model	Emotion model	Dataset
[6]	2014	Audio	LSTM, SVM	VA model	MediaEval Emotion
[5]	2016	Audio	DBLSTM	VA model	MediaEval Emotion
[34]	2017	Audio	CNN	18 classes	CAL500, CAL500exp
[50]	2017	Audio	LSTM, Attention	VA model	MediaEval Emotion
[36]	2018	Audio	BiLSTM	VA model	DEAM
[54]	2018	Audio	GARM	9 classes	Emotify
[35]	2019	EEG	CNN	VA model	Self-built
[46]	2019	Audio	BCRSN	VA model	DEAM, MTurk
[37]	2019	Audio	VGG	8 classes	Mid-level Perceptual Features, Soundtracks
[41]	2020	Audio	CNN	VA model	EmoMusic
[49]	2020	Audio	Attentive LSTM	VA model	MediaEval Emotion
[47]	2021	Audio	CLDNN	VA model	AnnoEmo
[51]	2021	Lyrics	XLNet	VA model	MoodyLyrics, MER Dataset
[42]	2022	Audio	CNN	VA model	DEAM
[52]	2022	Symbolic	MT-SMNN	VA model, 4 classes	VGMIDI, EMOPIA
[53]	2023	Audio	MCAN	VA model	DEAM, PMEmo
[48]	2023	Audio	Hybrid CNN-LSTM	9 classes	Emotify
[44]	2023	MIDI	RGRU-ARFA	4 classes	EMOPIA
[55]	2024	Audio	CL-RL model	4 classes	EMOPIA
[58]	2024	Audio	CBA	VA model	VioMusic
[56]	2024	Audio	MS-SincResNet, MS-SSincResNet	VA model	DEAM, PMEmo
[45]	2024	Audio	RNN, BRNN, LSTM	VA model	4Q audio, MTG-Jamendo
[57]	2024	Audio	Multi-scale Parallel Convolution	8 classes	PMEmo, Soundtrack, RAVDESS

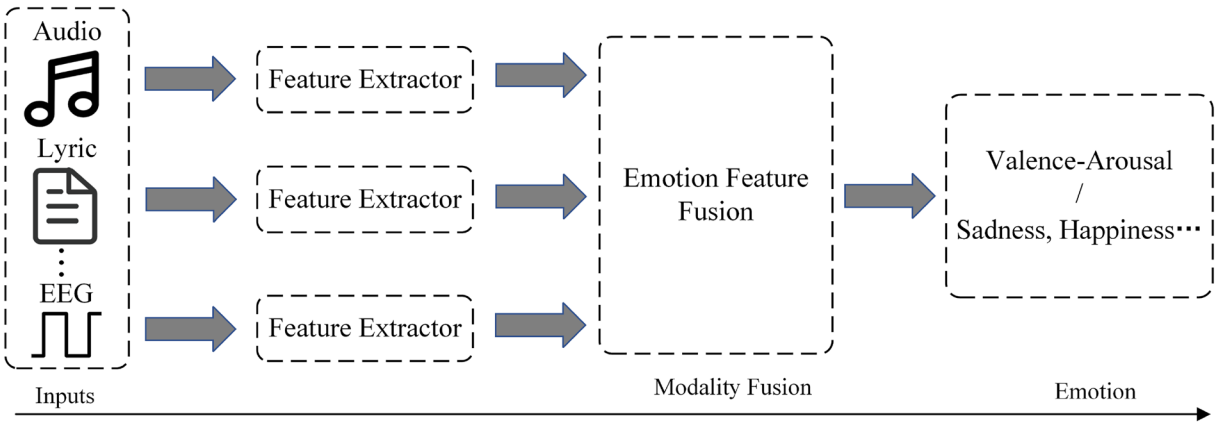

Fig. 6 The basic framework of multimodal deep learnig

for lyrics data. An additional layer of units connected the two DBMs to form a single model. Eventually, the features learned by the model were fed back to the SVM to predict the outcome. They also compared their proposed model with unimodal, early fusion, and late fusion, and experiments showed that their multimodal model outperformed the others, confirming the validity of multimodality. Furthermore, Silva et al. [39] proposed structuring musical features over a heterogeneous network and using the graph convolutional neural network to learn multi-modal features from audio and lyrics. Experiments show that the

heterogeneous graph neural network classifier exhibits superior performance in music emotion recognition.

Additionally, Delbouys et al. [60] tested various unimodal as well as multimodal models. The experimental results showed that for audio data, two layers of CNN processing get better results. For lyrics data, one layer of CNN plus LSTM can achieve better performance. In Zhou et al. 's work [61], they introduced the Bimodal Deep Auto Encoder model, which extracts features from both audio and lyrics to explore the correlation between these two modalities in music emotion regression. The results of experiments validated that unsupervised deep learning can effectively establish the relationship between two modes of data. In addition, through unimodal enhancement experiments, they discovered that the proposed deep network can learn better features in one modality (e.g., lyrics) if another modality (e.g., audio) is present at the same time as unsupervised feature learning. Chen et al. [62] presented a CNN-LSTM-based multi-feature combination network classifier for bridging the shortcomings of single-feature models. The classifier combined 2D audio features and 1D text features through the CNN-LSTM network to improve the classification performance. For audio feature extraction, they performed fine segmentation and vocal separation, then extracted spectrograms and low-level descriptors as feature representations. In lyrics feature extraction, the chi-square test vectors and word embeddings extracted by Word2vec were used as feature representations of the lyrics. Similarly, Sams et al. [63] designed a multimodal emotion recognition system. The audio data was represented by a mel spectrogram and CNN-LSTM was used to extract audio features, while the features of the lyrics data were extracted by XLNet. Then they were combined and used to train the ANN model by stacking integration method to obtain the best probabilistic weight output. In addition, Zhao et al. [38] proposed a multimodal multifaceted MER model based on professional knowledge of music psychology. The model extracts musically meaningful symbols and acoustic features from MIDI and audio data, such as the rhythm, dynamics, melody, harmony, and timbre of piano music, and fuses all information through an attention mechanism. Finally, they validated the performance of the model based on the EMOPIA dataset.

Moreover, knowledge distillation and transfer learning techniques have also been explored in the field of MER. Tong et al. [64] suggested a multimodal approach that combines knowledge distillation with music-style transfer learning. It is compared to traditional methods such as single audio, single lyrics, and single audio plus lyrics. The experimental results demonstrate that the proposed method improves emotion recognition accuracy and gets superior generalization ability. In addition, Shi et al. [65] were the first to propose combining self-supervised representation with handcrafted features inspired by music theory to capture emotional information. They designed a multimodal fusion module to comprehensively integrate the specific and invariant emotional information from speech and text modalities, using the emotional clues in different modalities to ensure a comprehensive understanding of cross-source emotional content. Wang et al. [66] proposed a Multilayered Music Decomposition and Multimodal Integration Interaction (MMD-MII) model, which constructs a new multimodal interaction framework to extract the current emotion vector at each time step, and further fuses and updates these emotion vectors to ensure the emotional consistency between various modalities. They also introduced a hierarchical framework based on music theory, focusing on the lead and chorus parts, and the chorus is processed separately to extract accurate emotional representation.

Some researchers have also turned their attention to particular forms of data. Such as, Jia [67] introduced an explicit sparse attention network based on CNN-LSTM. On the one hand, the model took audio features as network inputs, CNN-LSTM for spectrogram feature extraction and selection, and an explicit sparse attention network for output. On the other hand, the artificially designed low-level features LLD are combined into high-level statistical features (HSF) through statistical methods and then passed through DNN to extract features. Finally, the features of both networks were combined into audio fusion features, which were inputted to the Softmax layer for classification prediction. The experiment proved that the model was effective in recognizing and classifying music emotions. Also, the multimodal emotion recognition method proposed by Guo et at. [68], namely CSDAMER, is a CNN-SVM and data augmentation-based method. The method used physiological information such as electrocardiogram (ECG), galvanic skin response (GSR), and respiration (RSP) as inputs. Initially, they extracted high-level features of physiological signals by the convolutional layer of CNN. Then the extracted features were fed into SVM to obtain the results of emotion recognition. Experiments showed that CSDAMER achieved excellent results in both arousal and valence. Moreover, the use of data augmentation led to a further increase in the accuracy of arousal and valence. In addition to taking audio and lyrics as input information, Zhao et al. [69] also considered inputting the track name and artist as background information of the song into the model. The performance of the model was verified on MSDD and MoodyLyrics data sets. Zhu et al. [40] used a convolution-based acoustic encoder and a Transformer-based symbolic encoder to encode mixed acoustic features and musical score sequences respectively. Then, an attention mechanism was used to complete the interaction between modalities.

Generally, the application of multimodal technology has enabled MER research to focus not only on the improvement of unimodal data and networks, but also on the utilization

of complementarities between multimodal data to improve the performance of models. In addition, it makes up for the shortcomings of unimodal techniques that fail to extract diverse information. Meanwhile, the fused features can capture the emotional information conveyed by music more effectively. Table 6 presents the relevant work of multimodal deep learning.

4.3 Summary of learning models

Both traditional machine learning methods and deep learning methods have achieved outstanding results in music emotion recognition models, but both have their advantages and disadvantages.

Traditional machine learning methods rely on artificially designed features and shallow models. Firstly, audio features (such as MFCC, spectral features, rhythm features, etc.) are extracted through music signal processing technology, and then SVM, random forest, naive Bayes, and other algorithms are used for emotion classification. This structure gives it the advantages of high feature controllability, low computing resource requirements, stable performance of small data, and strong interpretability. However, the strong dependence on feature engineering, the difficulty in processing high-dimensional and unstructured music data make it difficult to cope with complex emotional changes in music.

Deep learning methods automatically learn multi-level features directly from raw audio or spectrograms by using models such as CNN, LSTM, and Transformer without manual intervention. Compared with machine learning methods, deep learning methods have the advantages of automatic learning features, a strong ability to process high-dimensional data and strong scalability. At the same time, it has the disadvantages of high data requirements, high computing resource requirements, and poor interpretability.

In general, traditional machine learning methods may be better in scenarios with limited resources and small amounts of data, but when pursuing high precision and complex tasks, deep learning is the inevitable choice.

5 Datasets and evaluation

In this section, we will introduce data sets and metrics commonly used in the MER field.

5.1 Datasets

Music emotion recognition necessitates rich, high-quality music data. With the rapid development of the internet in recent years, it has become effortless to acquire music resources and generate comprehensive datasets through classification and labeling. However, due to the subjectivity of MER research, the collection, labeling, and evaluation of datasets are the most challenging aspects. Currently, researchers in the MER field usually use self-built datasets for experiments, and some of these datasets can not be released due to copyright restrictions on music. In short, MER datasets can be categorized into two types: self-built datasets and public datasets.

In Table 7, some of the most frequently utilized public datasets are included in the list. The CAL500 [70] is a dataset comprising 502 Western popular music songs, used to evaluate music information retrieval systems. Each song in the CAL500 dataset contains audio data and multiple annotations covering instrumentation, vocal characteristics, genre, mood, song concepts, and use of terminology. Then, the CAL500exp [71] dataset is an extended version of the CAL500 music information retrieval dataset intended to enable Its aim is to aid in the automatic tagging of music

Table 6 MER methods based on the multimodal deep learning

Reference	Year	Feature modalities	Learning model	Emotion model	Dataset
[59]	2016	Audio, Lyric	DBM	4 classes	MSD
[60]	2018	Audio, Lyric	CNN, RNN, etc	VA model	MSD
[61]	2019	Audio, Lyric	BDAE	3 dimensions	Self-built
[62]	2020	Audio, Lyric	CNN-LSTM	5 classes	Million song dataset
[64]	2022	Audio, Lyric	CNN, LSTM	5 classes	Self-built
[68]	2022	ECG, GSR, RSP	CSDAMER	VA model	Self-built
[67]	2022	Audio, LLDs	CNN-LSTM	4 classes	Self-built
[69]	2022	Audio, Lyric	CNN, BERT	VA model	MSDD, MoodyLyrics
[39]	2022	Audio, Lyric	GCN	VA model	PMEmo
[38]	2023	Audio, Symbolic	CNN, Bi-GRU	VA model, 4 classes	EMOPIA
[63]	2023	Audio, Lyric	CNN-LSTM, XLNet	3 classes	Indonesian songs
[40]	2023	Audio, Symbolic	CNN, Transformer	4 classes, VA model	EMOPIA
[65]	2024	Audio, Lyric	WavLM, RoBERTa	4 classes	IEMOCAP
[66]	2024	Audio, Lyric	VGGish, ALBERT	4 classes	DEAM, FMA

Table 7 A summary of datasets

Dataset name	Emotion conceptualization	Number of songs	Data type	Genres
CAL500 [70]	Categorical	500	MP3	–
CAL500exp [71]	Categorical	3223 (segments)	MP3	–
AMG1608 [72]	Dimensional	1608	WAV	Rock, metal, country, jazz, etc
DEAM [73]	Dimensional	1802	MP3	Rock, pop, electronic, etc
Emotify [74]	Categorical	400	MP3	Rrock, classical, pop, electronic
EMOPIA [75]	Categorical	387 (1087 clips)	–	–
EmoMusic [76]	Dimensional	744	MP3	–
DEAP [77]	Dimensional	120	MP4	–

on a smaller time scale. CAL500exp tags, in contrast to CAL500, are annotated at the fragment level rather than the track level. Furthermore, the AMG1608 [72] dataset comprises frame-level acoustic characteristics obtained from 1,608 music clips lasting 30 s each, along with emotional valence-arousal (VA) annotations submitted by 665 participants. Additionally, the DEAM [73] dataset collects 1,802 pieces of music and is annotated by valence-arousal (VA). The Emotify [74] consists of 400 song excerpts (1 min long), including rock, classical, pop, and electronic. The EMOPIA [75] dataset is a multi-modal database, consisting of both audio and MIDI formats, that offers insight into perceived emotions in pop piano music, thus facilitating research on numerous music emotion tasks. The dataset comprises 1,087 music clips extracted from 387 songs and features clip-level emotion labels that have been annotated by four dedicated annotators. The EmoMusic [76] dataset is a dataset of 744 songs extracted from a dataset of 1000 songs selected from the Free Music Archive (FMA). It collects continuous VA ratings every second, with valence representing positive and negative emotions and arousal representing the intensity of emotions. The dataset is divided into two parts, one for the training set of 619 songs and another for the testing set of 125 songs. The DEAP [77] dataset is a multimodal dataset based on the analysis of human emotional states. The dataset consists of two parts, one is the ratings from online self-assessment, where 120 one-minute music video excerpts are rated by 14–16 volunteers based on arousal, valence, and dominance. The other part is the physiological cues from videos of 32 participants. Each participant watched 40 music videos and rated the emotional response based on video arousal, valence, dominance, and perception, and based on preference and familiarity, of which 22 participants' HIA recorded frontal facial videos.

However, there are still some shortcomings in the datasets. Firstly, a large part of the research in the MER field is carried out on privately constructed datasets, which can not be made public due to copyright issues. Secondly, most publicly available data usually contain hundreds of songs and few contain more than 2,000 songs, as shown in Table 7. At the same time, these public datasets also suffer from problems such as unbalanced distribution of song types and worrying label quality. The limitations mentioned above hinder the progress of researchers in the exploration of the MER field.

In view of the shortcomings of datasets in the field of MER, this paper provides the following suggestions to promote the development of MER. Firstly, regarding the issue of the use of copyright in music works, researchers can actively seek partnerships with music rights holders, streaming platforms, and independent artists to license the legal use of music works. Additionally, AI technologies such as GANs (Generative Adversarial Networks), VAEs (variational Auto-Encoders), etc. can be used to generate synthetic music samples that can be used for training and testing without worrying about copyright issues. Secondly, for the issue of small data sets, researchers can consider integrating multiple information sources such as audio, lyrics, text, and video content to build large multimodal data sets. Meanwhile, it is also possible to use crowdsourcing combined with automated data annotation tools to rapidly expand the size of datasets. Thirdly, for the issue of the unbalanced distribution, researchers can consider expanding existing data sets, analyzing the distribution of song types in the existing dataset, and adding new songs to balance the distribution of the dataset. In addition, the balance of data distribution is also considered when constructing new datasets. Finally, we discussed the issue of poor label quality. When constructing a new dataset, consider bringing in professional music critics or psychology experts to review the song selection and emotional labels to ensure data quality and consistency. Moreover, the large language models are considered for conducting a secondary audit to ensure the accuracy and detail of the label.

5.2 Evaluation metrics

For the MER model, objective evaluation metrics are necessary to assess its performance. In the field of MER,

commonly used evaluation metrics can be classified into two types: metrics for classification and regression models.

5.2.1 Classification evaluation metrics

For classification models, accuracy is one of the most widely used evaluation indicators. It evaluates the performance of the model by calculating the proportion of the number of correctly classified to the total number.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where, TP is true positive, FP is false positive, FN is false negative, TN is true negative, and the value of Accuracy ranges from 0 to 1. However, when facing positive and negative sample imbalances, the Accuracy has shortcomings in reflecting the model's predictive ability of a few categories of samples.

The F1-score is a metric that is used to evaluate the performance of a classification model. It is the coordinated average of precision and recall, which can evaluate the performance of the model more comprehensively.

$$\text{F1-score} = 2 \times \frac{TP}{2 \times TP + FP + FN} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

The definition of conformity is the same as in the above accuracy. The F1 score ranges from 0 to 1, and the higher the value, the better the model performance.

5.2.2 Regression evaluation metrics

R^2 is one of the most common model evaluation indicators in music emotion recognition, and is often used to evaluate regression models. R2 plays a key role in measuring the degree of fit between the regression model and the accuracy of the predictions. The main calculation formulas are as follows.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

It measures the performance of the regression model by comparing the sum of squares of residuals. R^2 determines the value range of the coefficients is $0 \sim 1$, and $R^2 = 1$ means that the model and the data are better fitted.

The root mean square error RMSE is the square root of the mean of the sum of squares of the difference between the observed value and the true value. It can effectively reflect the overall degree of deviation between the observed value and the true value, and is more sensitive to larger errors.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

where, the smaller the value of RMSE, the higher the prediction accuracy of the model.

The consistency correlation coefficient (CCC) is a statistical indicator that is used to evaluate the consistency between two continuous variables. It not only measures the correlation between variables, but also takes into account the absolute deviation between variables. CCC combines the characteristics of the Pearson correlation coefficient (PCC) and the mean square error (MSE) to measure the correlation and absolute consistency of the data at the same time.

$$CCC(x, y) = \frac{PCC(x, y) \cdot \sigma_x \cdot \sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x, \mu_y)^2} \tag{7}$$

$$CCCLoss = 1 - CCC(x, y) \tag{8}$$

where, the range of the value of CCC is between -1 and 1. And the closer the value of CCC is to 1, the stronger the consistency between the two variables. The value of CCCLoss varies between 0 and 2.

5.3 Performance and analysis on different datasets

As shown in Table 8, we show the performance of some models on the MSD, MediaEval Emotion, CAL500, EMOPIA, and MoodyLyrics datasets. It can be clearly observed from Table 8 that unimodal models have achieved outstanding results in the field of music emotion recognition. From the performance of [34] and [33] on the CAL500 dataset, it can be seen that using different signals as input has an impact on the performance of the model. In addition, comparing the results of [51] and [78] on the MoodyLyrics dataset shows that the multi-modal model has more outstanding performance compared to the uni-modal model. Multimodal models can use the complementarity between multimodal data to achieve better performance. Audio features can capture the emotional colors conveyed by musical elements such as melody and rhythm; while lyrics directly express the emotional content of the creator's intention. The combination of the two can compensate for the shortcomings or deviations that may exist in a single modality.

Although these methods have achieved great results, there are still some challenges. With the development of the field of artificial intelligence, large models are constantly being

Table 8 The performance of some methods on different datasets

References	Feature modalities	Learning model	Dataset	Performance
[59]	Audio, Lyrics	DBM	MSD	78.5% (Accuracy)
[60]	Audio, Lyrics	CNN, RNN, etc	MSD	0.219 for valence, 0.232 for arousal (R2)
[50]	Audio	LSTM, Attention	MediaEval Emotion	0.291 for valence, 0.241 for arousal (RMSE)
[43]	Audio	DBLSTM	MediaEval Emotion	0.285 for valence, 0.225 for arousal (RMSE)
[34]	EEG	CNN	CAL500	42.6% (Marco average precision)
[33]	Audio	CLR	CAL500	48.8% (Marco average precision)
[38]	Audio, Symbolic	CNN, Bi-GRU	EMOPIA	0.869 for valence, 0.884 for arousal
[40]	Audio, Symbolic	CNN, Transformer	EMOPIA	0.869 for valence, 0.874 for arousal
[51]	Lyrics	XLNet	MoodyLyrics	94.8% (F1-score)
[78]	Audio, Lyrics	CNN, BERT	MoodyLyrics	96.1% (F1-score)

proposed, such as ViT [79], CLIP [80], LLaMA [81], etc., and at the same time, higher quality and larger datasets are required for better performance.

6 Case studies and application analysis

MER is a multidisciplinary research field that involves psychology, medicine, and computer science. This field classifies the emotional state conveyed by music through analyzing music signals, features, or related data, and promotes the exploration of related fields.

6.1 Case studies

MER holds significant potential and demonstrates extensive application value across various domains. Firstly, MER offers robust technical support for personalized music recommendation systems [82–84]. By analyzing users' listening histories and preferences, these systems can recommend music that aligns with individual tastes. The integration of MER technology enhances the personalization and accuracy of such systems by identifying the emotional content of music and recommending tracks that resonate with the user's current mood. Moreover, MER has facilitated advancements in psychotherapy [85–88]. Music is extensively utilized in therapeutic settings to assist patients in expressing and managing their emotions. MER technology enables real-time monitoring of patient's emotional responses to music, allowing for dynamic adjustments in music selection, thereby enhancing the effectiveness of therapy sessions. Additionally, MER holds considerable promise in the field of neuroscience [89–92]. As an expressive tool, music elicits a wide range of emotional responses from listeners. MER provides the necessary technical framework to investigate the neural activity triggered by musical stimuli. Furthermore, MER demonstrates significant potential in diverse domains, including background music generation, smart home applications, and so on.

6.2 Application in psychotherapy

MER plays a pivotal role in the field of psychotherapy. Music therapy is a specialized form of psychotherapy that utilizes music and its evocation of human emotions to promote physical and mental well-being. Its origins can be traced back to ancient civilizations such as Greece and Rome [93]. The establishment of the American Music Therapy Association in 1944 marked the formal recognition of music therapy as a legitimate and regulated therapeutic method [94]. Since then, numerous researchers have delved into this field and achieved promising results.

Fachner et al. [95] conducted a study investigating the effects of music on anterior temporal lobe activity in resting patients with depression using an EEG-based music emotion recognition method to identify emotional changes induced by music. Additionally, Ramírez et al. [86] utilized EEG signals to evaluate the emotional responses of advanced cancer patients to music, thereby demonstrating the positive impact of music psychotherapy on patient emotions. Furthermore, Byrns et al. [87] proposed an EEG-based emotion recognition method to analyze how music influences the brain system, induces positive emotions, and mitigates negative emotions, thus alleviating the mental and cognitive symptoms associated with Alzheimer's disease. Ferreira et al. [88] introduced a generative mLSTM model capable of generating symbolic music based on specified emotions, enabling music researchers to tailor music for various application scenarios. For psychological researchers, automatic music composition offers a customized music therapy process for individuals with psychological disorders. Li et al. [85] designed a music recommendation system based on psychotherapy principles. This system selects the most appropriate music based on the user's past and current emotional states and

refines recommendations through a caring factor to enhance the user's mental health.

6.3 Application in neuroscience

Neuroscience is the study of the structure, function, development, genetics, and physiology of the nervous system. It combines multiple disciplines such as biology, psychology, medicine, and computer science to understand the brain and its impact on behavior, perception, emotion, and cognition. Exploring how music stimulates neural activity in the brain is already a recognized study field [93].

Since 1992, Steinberg et al. [96] through EEG signal analysis, found that different types of music have various effects on the EEG signal intensity in different frequency bands, and the interest in music-induced neural activity has continued to grow. Researchers Banerjee et al. [97] studied the effects of Hindustani music on brain activity in a relaxed state and found that the listeners' arousal levels were significantly improved under the inducement of music. In addition, Sanyal et al. [89] explored the reverse inversion of brain sounds, aiming to find correlations between EEG signals and musical stimuli, trying to understand how this correlation changes under the influence of different types of musical stimuli. Some studies generate emotion-related music by providing EEG signals. Considering that some fragments in a piece of music are unrelated to emotions, Li et al. [90] built a personalized reconfigurable music-EEG library to select music fragments for reconstruction and combination for different emotion regulation goals. Qiao et al. [91] proposed a CNN-SA-BiLSTM and explored the impact of different EEG frequency bands and multi-band combinations on music emotion recognition. The results showed that the emotional expression of the α band under music stimulation was particularly prominent, consistent with the results of related neuroscience research. Deng et al. [92] introduced a two-level feature selection method based on Support Vector Machine Recursive Feature Elimination (SVM-RFE) and a Random Forest (RF) algorithm to investigate the neural regulation behind painful emotions in patients with depression.

6.4 Application in daily life

MER research has a wide range of application scenarios. In clinical medicine, MER contributes to developing targeted treatments for mental health, cognitive impairment, stress management, depression, and anxiety. In neuroscience, MER provides new research ideas for studying music-induced neural activity. In addition, MER also has a wide range of application scenarios in daily life.

In daily life, MER research can promote progress in the field of music recommendation systems, improve user experience, and provide more personalized music recommendations. In addition, for some media communication platforms such as Weibo, the emotion recognition music recommendation system can recommend matching music based on the information content shared by users on the platform, making the platform more personalized. Deng et al. [83] believed that a good music recommendation system should be personalized and context-aware. Therefore, they proposed a sentiment-aware recommendation system that makes music recommendations based on the user's music library and the sentiment extracted from microblogs. Niu et al. [84] analyzed the current status of music classification and recommendation and found that the recommendation system that classifies music according to language, music style, theme scene, and time sequence fails to meet the requirements of accurate recommendation. Therefore, he proposed a music emotion recognition model based on gated recursive unit networks and multi-feature extraction, combined with the music recommendation model framework for topic classification, to achieve more accurate music recommendation. Lucia-Mulasd et al. [98]considered using music emotion recognition to facilitate automatic subtitle generation in videos. They combined the constant Q transform spectrogram with a convolutional neural network to propose an effective emotion classification model that helps automatically identify the emotional intent of different clips in movie soundtracks. Matoset et al. [99] proposed a hyperlapse method based on the emotional alignment of videos and songs. This method speeds up long videos while maintaining visual stability and matching the emotions conveyed by background music. This method finds the best matching path between video and music by selecting the optimal path, so that the emotions in the video are better combined with the background music, enhancing the audience's emotional experience.

For the field of psychotherapy, the emergence and development of MER provides a different method for the treatment of patients with mental illness. In terms of neuroscience, MER provides a research method to reveal the neural mechanism of emotion processing. In addition, the research of MER not only promotes the progress of the scientific field but also helps people enjoy life better and improve the quality of life. In short, MER plays an important role in psychotherapy, neuroscience, daily life, and other fields.

6.5 Summary of potential impacts of music emotion recognition

There are still some potential impacts on the application of music emotion recognition in daily life. On the one hand, the application of MER technology will prompt the music industry to innovate in the creation, dissemination, and consumption. Music creators can create music works that meet market demand and audience emotional preferences more

targetedly based on the results of sentiment analysis; music platforms can attract more user attention and participation through sentiment recommendation and sentiment marketing, and expand the music market; users can more conveniently obtain music that meets their emotional needs and improve their music consumption experience.

On the other hand, the combination of music emotion recognition and emotional computing will further promote the cross-integration of music with multiple disciplines such as psychology, computer science, and artificial intelligence, and will give birth to more emerging research directions and application scenarios, such as intelligent music psychotherapy, emotion-assisted medical diagnosis, etc., to provide new ideas and methods for solving problems related to human emotions.

However, with the widespread application of MER technology, it has also triggered some ethical and social issues. For example, over-reliance on emotional data for recommendation may cause users to be bound by "emotional cocoons", limiting users' exposure to and appreciation of different types of music; how to protect user privacy and data security in the process of collecting and using emotional data is also an urgent problem to be solved.

In general, MER has huge application potential and wide impact. It can bring more convenience and a better experience to people's lives. However, some ethical and social issues need to be addressed.

7 Future research directions

With the development of traditional machine learning and deep learning, artificial intelligence has made significant progress in the field of artificial intelligence. However, challenges remain, and many new trends have emerged. In this subsection, we will focus on challenges and future trends in the MER field.

7.1 Future research challenges

Firstly, we discuss several challenging issues in the field of MER and summarize them into the following four points:

(1) Subjectivity and individual differences. Musical emotions are highly subjective and difficult to quantify. Therefore, different people may show different emotional responses to the same piece of music, and even the same person may show different emotions at different times and situations. This is due to differences in individual cultural background, psychological state, emotional preferences, and other factors. For dimensional emotion modeling, arousal and valence models are usually used, each quadrant of which contains rich information about approximate emotions, but it is difficult to determine which emotion each value corresponds to, and it cannot be directly mapped to a specific emotion category. For example, "anger" and "excitement" may be located in the same area with high arousal but different valence, resulting in ambiguity in the classification results. For the classification emotion model, this simplified classification may not be sufficient to capture complex emotional states, which may cause the model to perform poorly when dealing with nuances. Therefore, the solution to this challenge is to further study new emotion models that can more accurately reflect individual emotions.

(2) Multi-modal data fusion. Music emotion recognition involves various modal data, such as audio, lyrics, sheet music, and so on. Effectively fusing information from different modalities is a challenging problem. The correlations and weight assignments between different modalities are crucial for the fusion of information. For example, the emotion words in lyrics are sparse, and sometimes the word order is changed for rhyming. Therefore, it is essential to explore appropriate feature extraction methods, feature fusion strategies, and model structures for the effective fusion of multimodal data.

(3) Lack of data and difficulty in labeling. The MER field needs some authoritative, large-scale, and diverse music emotion datasets. Currently, the public datasets are relatively limited, and the labeling often has subjective and inconsistent problems. The lack of sufficient data and high-quality labeled data limits the training and evaluation performance of the algorithm. For example, the number of songs in most of the data in Table 7 is less than 2,000, which is prone to overfitting for large models. In addition, there is a lack of authoritative datasets, so the models trained based on these limited data may have significant cultural biases and it is difficult to accurately identify the emotional content conveyed by music works in other cultural backgrounds.

(4) Feature extraction and modeling. Accurately extracting and modeling features of musical emotions is important and challenging. Musical emotion is diverse and dynamic. Traditional emotion representation methods, (e.g., MFCC, audio energy, bag-of-words modeling, etc.) may not adequately capture the complexity involved. Although deep learning models (e.g., RNN, Transformer, multimodal approaches, etc) provide us with powerful tools to better capture emotional information, further research is still needed on how to select appropriate features and how to design more effective networks.

7.2 Methods for challenges

Then, to address the challenges mentioned above, we propose some possible solutions as follows:

(1) Subjectivity and individual variability. To solve this problem, we need to study emotional models that better align with people's emotional changes. Based on a deep understanding of the mechanisms and expressions of human emotions, it is recommended to collaborate with psychologists and neuroscientists to conduct experiments on how music affects brain activity and mental states, and to apply these findings to the design of emotional models. Furthermore, considering that people's emotional responses to music may vary significantly across different cultural and social contexts, it is important to incorporate research findings from cultural and sociological studies to ensure that the emotional description model can adapt to a diverse global audience.
(2) Multi-modal data fusion. Consider using advanced feature extraction techniques to extract features from different modalities, the quality of feature extraction directly affects the performance of MER models. For audio, open-source pre-trained models such as VGGish, AST, etc. can be considered for use to extract features. For lyrics, pre-trained language models such as BERT can capture the semantic information in the text. Obviously, feature fusion can capture the complementarity between different modal information and enhance semantic expression. Therefore, we recommend designing an adaptive mechanism to adjust the relative importance of different modalities according to the specific situation. For example, when there are no lyrics, it relies more on audio features. When there are clear lyrics, the proportion of text information is increased.
(3) Lack of data and difficulty in labeling. For this issue, AI technologies such as GANs and VAEs can be used to generate synthetic music samples and their corresponding emotional labels to expand the amount of available data. These synthetic data can be used to supplement real data sets, especially when it is difficult to obtain copyright permission. Meanwhile, manual labeling combined with automation can also be used to deal with the issue of data labeling. On the one hand, volunteers are invited to participate in the labeling process. On the other hand, fine-tuned large language models or other machine learning algorithms are used to assist manual labeling to improve efficiency and reduce errors.
(4) Feature extraction and modeling. For audio signals, in addition to MFCC, we recommend exploring more advanced acoustic features, such as Chroma Features, Spectral Contrast, etc., to capture the emotional color of music. In addition, feature extraction can be performed on multiple time scales considering the characteristics of musical emotions changing over time. For example, short-term features are used to capture rapidly changing emotional fluctuations, while long-term features help reflect the overall emotional trend. Consider adopting a transfer learning strategy to pre-transfer relevant knowledge from related tasks or other fields, and then fine-tune it for the MER task.

7.3 Future research trends

Finally, we have investigated the development direction of MER field in recent years, and its future development can be summarized as follows:

(1) From traditional ML models to DL models. Compared to traditional ML techniques, DL techniques offer more significant advantages in the MER field due to their simplicity and effectiveness. Traditional ML methods require handcrafted features, while DL methods are able to automatically learn feature representations, simplifying the process of model learning and training. In addition, handcrafted features exhibit limited performance across various ML models and datasets.
(2) From static processing to dynamic processing. Static processing assigns an emotion label to an entire musical fragment, ignoring temporal changes in emotion and focusing instead on the overall emotional expression. Conversely, dynamic processing is concerned with temporal changes in emotional states, taking into account the temporal information in music. Therefore, dynamic processing is more in line with the characteristics of music, the emotion will change dynamically within a music piece. Especially, with the appearance of sequence models such as RNN, and LSTM, dynamic recognition of continuous emotions becomes easier and more accurate.
(3) From unimodality to multimodality. After years of development, methods based on unimodal techniques have hit a bottleneck, necessitating the exploration of alternative methods. Fortunately, the multimodal technology overcomes the bottleneck. By synthesizing information from multiple modalities, multimodal technology can compensate for the limitations of a single modality and improve the accuracy of emotion recognition. And fused features can more effectively capture the emotional information conveyed through music.
(4) From multimodal models to multimodal large language. With the proposed large language model, the research in the field of artificial intelligence has entered a new

era. ChatGPT, LLAMA, Qwen, and other large language models have strong context understanding and reasoning ability, and introducing them into the field of MER can effectively promote development. Moreover, the multimodal large language model can capture more detailed and complex emotion expression information when processing cross-modal information.

8 Conclusion

Music emotion recognition draws on traditional machine learning and deep learning algorithms to assign emotion labels to music as a whole or segments, allowing the listener to access the piece of music. Meanwhile, it also has a wide range of application scenarios in the fields of music recommendation systems [100], intelligent music creation and generation [101], automatic music annotation and indexing, and so on. This paper describes advances related to music emotion recognition in terms of feature extraction and emotion recognition. It first revolves around traditional machine learning, summarizing existing work from both hand-crafted features and machine learning models. Then, deep learning-based approaches are introduced, summarizing the existing related work from both unimodal and multimodal techniques, respectively. Next, some common datasets and rubrics in the field of MER are introduced. Finally, we discuss some application scenarios of the MER.

Overall, music emotion recognition can help people understand the music work more comprehensively, to dig deeper into its emotional connotation. Secondly, it can also provide technical support for music recommendations, personalized recommendations, the mental health field, etc. However, there is an urgent need for authoritative, large-scale, diverse datasets and more accurate emotion models in the MER field. How to design a reasonable network to extract complex features and how to utilize the extracted features effectively are the keys to MER research. Generally speaking, transferring from static process to dynamic process, from unimodal to multimodal, and from traditional ML model to DL model are effective measures to cope with the above problems. Therefore, the study of MER will be one of the indispensable research directions in the field of MIR.

Acknowledgements This paper is supported by the Sichuan Science and Technology Program (2024YFFK0120), the Major Project of Sichuan Provincial Natural Science Foundation (25ZNSFSC0022), and the National Science Foundation of China (62072419).

Author Contributions Y.W., X.Z. and Y.X. developed the framework of the manuscript and wrote the introduction part. Y.W., Y.X., H.R. and C.D. wrote the main parts of manuscript text, Y.X., X.Z., P.J. and X.C. revised and edited the paper and provided supervision for the entire work.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Cañón, J., Cano, E., Eerola, T., Herrera, P., Hu, X., Yang, Y., Gómez, E.: Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications. IEEE Signal Process. Mag. **38**, 106–114 (2021)
2. Cheng, Z., Shen, J., Zhu, L., Kankanhalli, M.S., Nie, L.: Exploiting music play sequence for music recommendation. In: Sierra, C. (ed.) IJCAI, pp. 3654–3660 (2017)
3. Yu, C., Wang, M.: Survey of emotion recognition methods using eeg information. Cognit. Robot. **2**, 132–146 (2022)
4. Yang, Y., Lin, Y., Su, Y., Chen, H.H.: A regression approach to music emotion recognition. IEEE Trans. Speech Audio Process. **16**(2), 448–457 (2008)
5. Li, X., Xianyu, H., Tian, J., Chen, W., Meng, F., Xu, M., Cai, L.: A deep bidirectional long short-term memory based multi-scale approach for music dynamic emotion prediction. In: ICASSP, pp. 544–548 (2016)
6. Weninger, F., Eyben, F., Schuller, B.W.: On-line continuous-time music mood regression with deep recurrent neural networks. In: ICASSP, pp. 5412–5416 (2014)
7. Hevner, K.: Experimental studies of the elements of expression in music. Am. J. Psychol. **48**, 246–268 (1936)
8. Posner, J., Russell, J.A., Peterson, B.S.: The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev. Psychopathol. **17**, 715–734 (2005)
9. Tellegen, A., Watson, D., Clark, L.A.: On the dimensional and hierarchical structure of affect. Psychol. Sci. **10**(4), 297–303 (1999)
10. Liu, X.-n.: A music emotion classifier construction algorithm of neural network based on relevant feedback. J. Northwest Univ. **71**(33), 267–282 (2012)
11. Wang, J., Yang, Y., Wang, H., Jeng, S.: The acoustic emotion gaussians model for emotion-based music annotation and retrieval. In: Babaguchi, N., Aizawa, K., Smith, J.R., Satoh, S., Plagemann, T., Hua, X., Yan, R. (eds.) ACM MM, pp. 89–98 (2012)
12. Wu, B., Zhong, E., Horner, A., Yang, Q.: Music emotion recognition by multi-label multi-layer multi-instance multi-view learning. In: Hua, K.A., Rui, Y., Steinmetz, R., Hanjalic, A., Natsev, A., Zhu, W. (eds.) ACM MM, pp. 117–126 (2014)
13. Xianyu, H., Li, X., Chen, W., Meng, F., Tian, J., Xu, M., Cai, L.: SVR based double-scale regression for dynamic emotion prediction in music. In: ICASSP, pp. 549–553 (2016)
14. Lartillot, O., Toiviainen, P.: MIR in matlab (II): A toolbox for musical feature extraction from audio. In: ISMIR, pp. 127–130 (2007)
15. Mathieu, B., Essid, S., Fillon, T., Prado, J., Richard, G.: Yaafe, an easy to use and efficient audio feature extraction software. In: ISMIR, pp. 441–446 (2010)
16. Wen, H.: Review on speech emotion recognition. J. Softw. **25**(1), 37–50 (2014)
17. Chen, P., Zhao, L., Xin, Z., Qiang, Y., Zhang, M., Li, T.: A scheme of MIDI music emotion classification based on fuzzy

theme extraction and neural network. In: CIS, pp. 323–326 (2016)
18. Barthet, M., Fazekas, G., Sandler, M.: Multidisciplinary perspectives on music emotion recognition: Implications for content and context-based models (2012)
19. Hu, X., Downie, J.S., Ehmann, A.F.: Lyric text mining in music mood classification. In: ISMIR, pp. 411–416 (2009)
20. Hu, Y., Chen, X., Yang, D.: Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In: ISMIR, pp. 123–128 (2009)
21. Dakshina, K., Sridhar, R.: LDA Based Emotion Recognition from Lyrics, pp. 187–194 (2014)
22. Thammasan, N., Fukui, K., Numao, M.: Application of deep belief networks in eeg-based dynamic music-emotion recognition. In: IJCNN, pp. 881–888 (2016)
23. Hu, X., Li, F., Ng, T.J.: On the relationships between music-induced emotion and physiological signals. In: ISMIR, pp. 362–369 (2018)
24. Nawa, N.E., Callan, D.E., Mokhtari, P., Ando, H., Iversen, J.R.: Decoding music-induced experienced emotions using functional magnetic resonance imaging - preliminary results. In: IJCNN, pp. 1–7 (2018)
25. Chen, Y.-A., Wang, J.-C., Yang, Y.-H., Chen, H.: Linear regression-based adaptation of music emotion recognition models for personalization. In: (ICASSP), pp. 2149–2153 (2014)
26. Chiang, W.C., Wang, J.S., Hsu, Y.L.: A music emotion recognition algorithm with hierarchical svm based classifiers. In: 2014 International Symposium on Computer, Consumer and Control, pp. 1249–1252 (2014)
27. Soleymani, M., Aljanaki, A., Yang, Y., Caro, M.N., Eyben, F., Markov, K., Schuller, B.W., Veltkamp, R.C., Weninger, F., Wiering, F.: Emotional analysis of music: A comparison of methods. In: ACM MM, pp. 1161–1164 (2014)
28. Fukayama, S., Goto, M.: Music emotion recognition with adaptive aggregation of gaussian process regressors. In: ICASSP, pp. 71–75 (2016)
29. Chen, Y., Wang, J., Yang, Y., Chen, H.H.: Component tying for mixture model adaptation in personalization of music emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing **25**, 1409–1420 (2017)
30. Markov, K., Iwata, M., Matsui, T.: Music emotion recognition using gaussian processes. In: Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, October 18-19, 2013, vol. 1043 (2013)
31. Malheiro, R., Panda, R., Gomes, P., Paiva, R.P.: Emotionally-relevant features for classification and regression of music lyrics. IEEE Trans. Affect. Comput. **9**, 240–254 (2018)
32. Chen, S., Lee, Y., Hsieh, W., Wang, J.: Music emotion recognition using deep gaussian process. In: APSIPA, pp. 495–498 (2015)
33. Liu, Y., Liu, Y., Zhao, Y., Hua, K.A.: What strikes the strings of your heart? - feature mining for music emotion analysis. IEEE Trans. Affect. Comput. **6**, 247–260 (2015)
34. Liu, X., Chen, Q., Wu, X., Liu, Y., Liu, Y.: CNN based music emotion classification. CoRR **abs/1704.05665** (2017)
35. Keelawat, P., Thammasan, N., Kijsirikul, B., Numao, M.: Subject-independent emotion recognition during music listening based on eeg using deep convolutional neural networks. In: CSPA, pp. 21–26 (2019)
36. Liu, H., Fang, Y., Huang, Q.: Music emotion recognition using a variant of recurrent neural network. In: Proceedings of the 2018 International Conference on Mathematics, Modeling, Simulation and Statistics Application (MMSSA 2018) (2019)
37. Chowdhury, S., Vall, A., Haunschmid, V., Widmer, G.: Towards explainable music emotion recognition: The route via mid-level features. In: ISMIR, pp. 237–243 (2019)
38. Zhao, J., Yoshii, K.: Multimodal multifaceted music emotion recognition based on self-attentive fusion of psychology-inspired symbolic and acoustic features. In: APSIPA ASC, pp. 1641–1645 (2023)
39. Silva, A.C.M., Silva, D.F., Marcacini, R.M.: Heterogeneous graph neural network for music emotion recognition. In: ISMIR, pp. 667–674 (2022)
40. Zhu, K., Zhang, X., Wang, J., Cheng, N., Xiao, J.: Symbolic and acoustic: Multi-domain music emotion modeling for instrumental music. In: ADMA, vol. 14179, pp. 168–181 (2023)
41. Yang, P., Kuang, S., Wu, C., Hsu, J.: Predicting music emotion by using convolutional neural network. In: HCIBGO, vol. 12204, pp. 266–275 (2020)
42. Orjesek, R., Jarina, R., Chmulik, M.: End-to-end music emotion variation detection using iteratively reconstructed deep features. Multimedia Tools and Applications **81**, 5017–5031 (2022)
43. Li, X., Tian, J., Xu, M., Ning, Y., Cai, L.: Dblstm-based multi-scale fusion for dynamic emotion prediction in music. In: ICME, pp. 1–6 (2016)
44. Kumar, V.B., Kathiravan, M.: Emotion recognition from midi musical file using enhanced residual gated recurrent unit architecture. Frontiers in Computer Science (2023)
45. Chang, X., Zhang, X., Zhang, H., Ran, Y.: Music emotion prediction using recurrent neural networks. ArXiv **abs/2405.06747** (2024)
46. Dong, Y., Yang, X., Zhao, X., Li, J.: Bidirectional convolutional recurrent sparse network (BCRSN): an efficient model for music emotion recognition. IEEE Trans. Multimedia **21**, 3150–3163 (2019)
47. Hizlisoy, S., Yildirim, S., Tufekci, Z.: Music emotion recognition using convolutional long short term memory deep neural networks. Engineering Science and Technology, an International Journal **24**, 760–767 (2021)
48. Yakovyna, V.S., Korniienko, V.V.: Music emotion classification using a hybrid cnn-lstm model. Applied Aspects of Information Technology (2023)
49. Chaki, S., Doshi, P., Patnaik, P., Bhattacharya, S.: Attentive rnns for continuous-time emotion prediction in music clips. In: AAAI, vol. 2614, pp. 36–46 (2020)
50. Ma, Y., Li, X., Xu, M., Jia, J., Cai, L.: Multi-scale context based attention for dynamic music emotion prediction. In: ACM MM, pp. 1443–1450 (2017)
51. Agrawal, Y., Shanker, R.G.R., Alluri, V.: Transformer-based approach towards music emotion recognition from lyrics. In: ECIR, vol. 12657, pp. 167–175 (2021)
52. Qiu, J., Chen, C.L.P., Zhang, T.: A novel multi-task learning method for symbolic music emotion recognition. CoRR **abs/2201.05782** (2022)
53. Zhang, M., Zhu, Y., Zhang, W., Zhu, Y., Feng, T.: Modularized composite attention network for continuous music emotion recognition. MMultimedia Tools and Applications **82**, 7319–7341 (2023)
54. Chang, W., Li, J., Lin, Y., Lee, C.: A genre-affect relationship network with task-specific uncertainty weighting for recognizing induced emotion in music. In: ICME, pp. 1–6 (2018)
55. Zhang, Y., Cai, D., Zhang, D.: Application and algorithm optimization of music emotion recognition in piano performance evaluation. Environ. Soc. Psychol. **9**(4), 1–16 (2024)
56. Chang, P.-C., Chen, Y.-S., Lee, C.-H.: Iiof: Intra- and inter-feature orthogonal fusion of local and global features for music emotion recognition. Pattern Recogn. **148**, 110200 (2024)
57. Han, X., Chen, F., Ban, J.: A gai-based multi-scale convolution and attention mechanism model for music emotion recognition and recommendation from physiological data. Appl. Soft Comput. **164**, 112034 (2024)

58. Ma, S., Zhou, R.: Violin music emotion recognition with fusion of cnn-bigru and attention mechanism. Inf. **15**(4), 224 (2024)
59. Huang, M., Rong, W., Arjannikov, T., Jiang, N., Xiong, Z.: Bi-modal deep boltzmann machine based musical emotion classification. In: ICANN, vol. 9887, pp. 199–207 (2016)
60. Delbouys, R., Hennequin, R., Piccoli, F., Royo-Letelier, J., Moussallam, M.: Music mood detection based on audio and lyrics with deep neural net. In: ISMIR, pp. 370–375 (2018)
61. Zhou, J., Chen, X., Yang, D.: Multimodel Music Emotion Recognition Using Unsupervised Deep Neural Networks, pp. 27–39 (2019)
62. Chen, C., Li, Q.: A multimodal music emotion classification method based on multifeature combined network classifier. Math. Probl. Eng. **2020**, 1–11 (2020)
63. Sams, A.S., Zahra, A.: Multimodal music emotion recognition in indonesian songs based on CNN-LSTM, XLNet transformers. Bulletin of Electrical Engineering and Informatics **12**, 355–364 (2023)
64. Tong, G., Ding, B.: Multimodal music emotion recognition method based on the combination of knowledge distillation and transfer learning **2022**, 13 (2022)
65. Shi, X., Li, X., Toda, T.: Multimodal fusion of music theory-inspired and self-supervised representations for improved emotion recognition. In: Annual Conference of the International Speech Communication Association, pp. 2024–2350 (2024)
66. Wang, J., Sharifi, A., Gadekallu, T.R., Shankar, A.: Mmd-mii model: A multilayered analysis and multimodal integration interaction approach revolutionizing music emotion classification. Int. J. Comput. Intell. Syst. **17**, 99 (2024)
67. Jia, X., Bhardwaj, A.: Music emotion classification method based on deep learning and explicit sparse attention network. Comput. Intell. Neurosci. **2022**, 9 (2022)
68. Guo, G., Gao, P., Zheng, X., Ji, C.: Multimodal emotion recognition using CNN-SVM with data augmentation. In: BIBM, pp. 3008–3014 (2022)
69. Zhao, J., Ru, G., Yu, Y., Wu, Y., Li, D., Li, W.: Multimodal music emotion recognition with hierarchical cross-modal attention network. In: ICME, pp. 1–6 (2022)
70. Turnbull, D., Barrington, L., Torres, D.A., Lanckriet, G.R.G.: Towards musical query-by-semantic-description using the CAL500 data set. In: Kraaij, W., Vries, A.P., Clarke, C.L.A., Fuhr, N., Kando, N. (eds.) SIGIR, pp. 439–446 (2007)
71. Wang, S., Wang, J., Yang, Y., Wang, H.: Towards time-varying music auto-tagging based on CAL500 expansion. In: ICME, pp. 1–6 (2014)
72. Chen, Y., Yang, Y., Wang, J., Chen, H.H.: The AMG1608 dataset for music emotion recognition. In: ICASSP, pp. 693–697 (2015)
73. Aljanaki, A., Yang, Y.-H., Soleymani, M.: Developing a benchmark for emotional analysis of music. PLoS ONE **12**(3), e0173392 (2017)
74. Eerola, T., Vuoskoski, J.K.: A comparison of the discrete and dimensional models of emotion in music. Psychol. Music **39**, 18–49 (2011)
75. Hung, H., Ching, J., Doh, S., Kim, N., Nam, J., Yang, Y.: EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation. In: ISMIR, pp. 318–325 (2021)
76. Soleymani, M., Caro, M.N., Schmidt, E.M., Sha, C.-Y., Yang, Y.-H.: 1000 songs for emotional analysis of music. In: CrowdMM '13 (2013)
77. Koelstra, S., Mühl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I.: Deap: A database for emotion analysis; using physiological signals. IEEE Trans. Affect. Comput. **3**, 18–31 (2012)
78. Zhao, J., Ru, G., Yu, Y., Wu, Y., Li, D., Li, W.: Multimodal music emotion recognition with hierarchical cross-modal attention network. In: ICME, pp. 1–6 (2022)
79. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
80. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) ICML. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763 (2021)
81. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. CoRR **abs/2302.13971** (2023)
82. Zhao, H., Jin, L.: Iot-based approach to multimodal music emotion recognition. Alex. Eng. J. **113**, 19–31 (2025)
83. Deng, S., Wang, D., Li, X., Xu, G.: Exploring user emotion in microblogs for music recommendation. Expert Syst. Appl. **42**, 9284–9293 (2015)
84. Niu, N.: Music emotion recognition model using gated recurrent unit networks and multi-feature extraction. Mobile Information Systems (2022)
85. Liu, Z., Xu, W., Zhang, W., Jiang, Q.: A music recommendation system based on psychotherapy. Science Talks **6**, 100222 (2023)
86. Ramírez, R., Planas, J., Escudé, N., Mercadé, J.J., Farriols, C.: Eeg-based analysis of the emotional effect of music therapy on palliative care cancer patients. Front. Psychol. **9**, 324998 (2018)
87. Byrns, A., Abdessalem, H., Cuesta, M., Bruneau, M., Belleville, S., Frasson, C.: Eeg analysis of the contribution of music therapy and virtual reality to the improvement of cognition in alzheimers disease. J. Biomed. Sci. Eng. **13**, 187–201 (2020)
88. Ferreira, L., Whitehead, J.: Learning to generate music with sentiment. In: Flexer, A., Peeters, G., Urbano, J., Volk, A. (eds.) ISMIR, pp. 384–390 (2019)
89. Nag, S., Sanyal, S., Banerjee, A., Sengupta, R., Ghosh, D.: Music of brain and music on brain: a novel eeg sonification approach. Cogn. Neurodyn. **13**, 13–31 (2017)
90. Li, Y., Zheng, W.: Emotion recognition and regulation based on stacked sparse auto-encoder network and personalized reconfigurable music. Mathematics **9**(6), 593 (2021)
91. Qiao, Y., Mu, J., Xie, J., Hu, B., Liu, G.: Music emotion recognition based on temporal convolutional attention network using EEG. Front. Hum. Neurosci. **18**, 1324897 (2024)
92. Deng, J., Chen, Y., Zeng, W., Luo, X., Li, Y.: Brain response of major depressive disorder patients to emotionally positive and negative music. J. Mol. Neurosci. **72**, 2094–2105 (2022)
93. Su, Y., Liu, Y., Xiao, Y., Ma, J., Li, D.: A review of artificial intelligence methods enabled music-evoked EEG emotion recognition and their applications. Front. Neurosci. **18**, 1400444 (2024)
94. Taylor, D.B.: Music in general hospital treatment from 1900 to 1950. J. Music Ther. **18**(2), 62–73 (1981)
95. Fachner, J., Gold, C., Erkkilä, J.: Music therapy modulates fronto-temporal activity in rest-eeg in depressed clients. Brain Topogr. **26**, 338–354 (2012)
96. Steinberg, R., Giinther, W., Stiltz, I., Rondot, P.: Eeg-mapping during music stimulation. Psychomusicology: A Journal of Research in Music Cognition **11**, 157–170 (1992)
97. Banerjee, A., Sanyal, S., Patranabis, A., Banerjee, K., Guhathakurta, T., Sengupta, R., Ghosh, D., Ghose, P.: Study on brain dynamics by non linear analysis of music induced eeg

signals. Physica A-statistical Mechanics and Its Applications **444**, 110–120 (2016)
98. Lucia-Mulas, M.J., Revuelta-Sanz, P., Ruíz-Mezcua, B., González-Carrasco, I.: Automatic music emotion classification model for movie soundtrack subtitling based on neuroscientific premises. Appl. Intell. **53**, 27096–27109 (2023)
99. Matos, D., Ramos, W., Silva, M., Romanhol, L., Nascimento, E.R.: A multimodal hyperlapse method based on video and songs' emotion alignment. Pattern Recognit. Lett. **166**, 174–181 (2022)
100. Deng, S., Wang, D., Li, X., Xu, G.: Exploring user emotion in microblogs for music recommendation. Expert Syst. Appl. **42**(23), 9284–9293 (2015)
101. Ferreira, L., Whitehead, J.: Learning to generate music with sentiment. In: ISMIR, pp. 384–390 (2019)