IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 August 2024, accepted 15 October 2024, date of publication 22 October 2024, date of current version 5 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3484470

TOPICAL REVIEW

Music Emotion Recognition Based on Deep Learning: A Review

XINGGUO JIANG[1,2], YUCHAO ZHANG[1], GUOJUN LIN[1,2], AND LING YU[1]
[1]School of Automation and Information Engineering, Sichuan University of Science and Engineering, Zigong 643000, China
[2]Artificial Intelligence Key Laboratory of Sichuan Province, Yibin 644005, China

Corresponding author: Yuchao Zhang (2300575448@qq.com)

This work was supported in part by the Scientific Research Foundation of Sichuan University of Science and Engineering under Grant 2019RC12, and in part by Sichuan Science and Technology Program under Grant 2024YFHZ0026.

ABSTRACT In recent years, with the development of the digital era, music emotion recognition technology has been widely used in the fields of music recommendation system, music classification, psychotherapy, music visualization, background music generation, smart home, and other applications of music emotion recognition, and has received attention from all walks of life. Especially the rapid development of artificial intelligence and deep learning, the music emotion recognition model using efficient deep neural network composition has become the mainstream model. This paper provides a more detailed overview of music emotion recognition, first introducing the background of music and emotion, and briefly summarizing the content of related works as well as the content framework. In the process, we also compare the similarities and differences in the content of other researchers' reviews of related research areas. And in the middle section, we provide a detailed account of datasets, emotion models, feature extraction, and emotion recognition algorithms. Finally, we discuss the current challenges in music emotion recognition and explore future research priorities.

INDEX TERMS Music emotion recognition, deep learning, artificial intelligence, music emotion datasets.

I. INTRODUCTION

Music is another vehicle for transferring emotions between people after speech. Perlovsky [1] argued that in primitive societies human vocalizations fell into two types: those that were less emotional but more semantically specific, and thus evolved into the language today; and those that retained an emotional connection with accompanying semantic ambiguity, and thus evolved into the music we have today. As a result, music tends to be more capable of conveying emotion and creating emotional resonance between the listener and the creator. Emotional expression in music has a certain universality, for example, opera in the West and drama in the East, both of which convey emotions vividly and imaginatively in the form of musical backgrounds and character performances. Music has a long history and is one of the most important representatives of human civilization. Western music can be traced back to the ancient Greek and Roman periods, and the ancient Greek philosophers believed that the human mind is very susceptible to negative emotions and that appropriate music can alleviate the effects of such negative emotions [1], which also confirms the feasibility of music therapy. Oriental music has a long history as well. According to the ancient civilizations studied, oriental music can be traced back to the Yellow River Valley of China in the early Neolithic period 8,000 years ago, and took shape in the late Neolithic period 5,000 years ago. Fang [2] believed that music civilization belongs to the category of spiritual civilization, music has become the emotional support, the story of ancient Chinese Bo Ya's string is a vivid example, and music civilization has become an important part of the Chinese civilization and the main spiritual mark.

Music emotion recognition is an interdisciplinary field of research that incorporates knowledge and techniques from disciplines such as musicology, psychology, computer science, and neuroscience. Music is a derivative of speech, and

The associate editor coordinating the review of this manuscript and approving it for publication was Zijian Zhang.

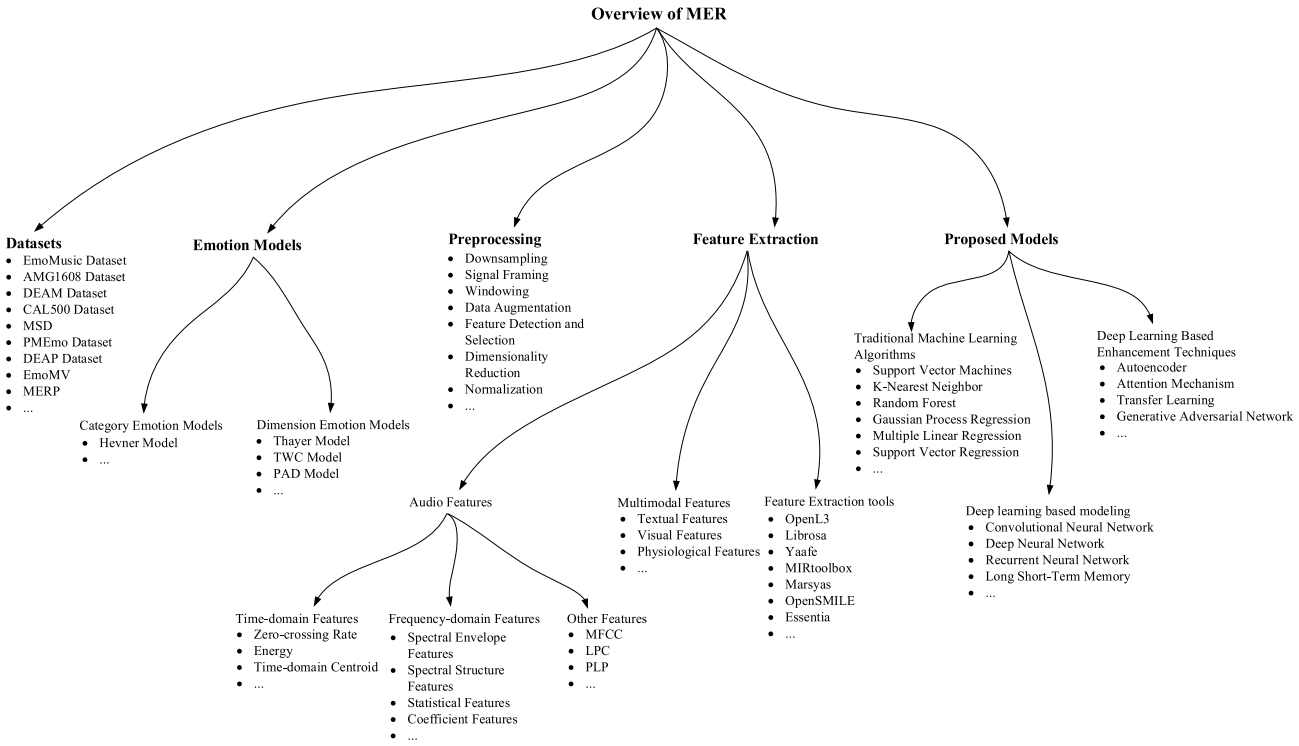


FIGURE 1. This is a diagram summarizing the content of our review of music emotion recognition (MER), and we have developed our review of music emotion recognition in order from left to right.

music emotion recognition has certain similarities to speech emotion recognition. But unlike speech emotion recognition, music emotion recognition includes more feature information such as melody, harmony, rhythm, timbre, and other musical features. Different styles of music and different musical instruments will also lead to differences in the expression of emotions. Coupled with challenges such as the lack of music emotion datasets, music emotion recognition is therefore more challenging than speech emotion recognition. In previous research, traditional machine learning algorithms such as support vector machines (SVM), random forests (RF), and k-nearest neighbors (K-NN) have been applied to music emotion recognition tasks. With the rapid development of artificial intelligence, deep learning has been widely used in music emotion recognition work, and most of the models proposed by researchers in recent years are based on the direct use or indirect evolution of convolutional neural networks (CNN) and recurrent neural networks (RNN). We have conducted a review based on the literature related to music emotion recognition in recent years, as shown in the general framework in Fig. 1. We focus on the last five years of research on deep learning models for music emotion recognition, and also provide a brief description of traditional machine learning models for comparison. The main contributions of this paper are as follows: (1) a dataset lacking in music emotion recognition is supplemented with a narrative summary of its content; (2) describe and categorize commonly used music emotion models and their development; (3) the scientific research results of deep learning models for music emotion recognition in the past five years are analyzed in detail, and the related applications of traditional machine learning algorithms are briefly described; (4) add a description of commonly used model evaluation methods and evaluation results of the proposed models, as well as examples of relevant applications in various fields of music emotion recognition; (5) present our perspectives on current challenges in music emotion recognition and our exploration of future research priorities.

II. RELATED WORK

We have researched previous literature reviews on music emotion recognition and listed their articles in Table 1 to facilitate comparison with the literature review we have described, and ''Partial'' in the table indicates that the relevant contents are small or missing. Due to the different focus of each researcher on music emotion recognition and the early publication of the review, there is incomplete or missing content in music emotion recognition. In Kim et al. [3], they focused on contextual textual information methods and content-based methods in music emotion recognition research, and investigated systems that combine multiple feature domains. Eerola et al. [4] focused on emotional modeling as well as music emotion recognition methods, although they and Yang et al. [5] only reported traditional machine learning models in their recognition methods due to the fact that deep learning was not widely used before 2012. The review by

TABLE 1. Comparison of the content of the literature review.

Article	Year	Dataset	Emotion Model	Feature	Preprocessing	Proposed Model	Model Evaluation
Music emotion recognition: A state of the art review [3]	2010	Partial	Partial	√	×	Partial	×
A review of music and emotion studies: Approaches, emotion models, and stimuli [4]	2012	×	√	√	×	Partial	×
Machine recognition of music emotion: A review [5]	2012	Partial	√	√	Partial	Partial	×
Review of data features-based music emotion recognition methods [6]	2018	Partial	Partial	√	Partial	Partial	×
Review of data features-based music emotion recognition methods [7]	2020	×	Partial	√	×	Partial	×
A survey of music emotion recognition [8]	2022	Partial	Partial	√	Partial	√	Partial
A review: Music-emotion recognition and analysis based on EEG signals [9]	2022	√	√	√	Partial	√	×
Our Research	2024	√	√	√	√	√	√

Yang et al. [6] focused on music emotion recognition methods for different data features of music features, ground truth data, and their combinations above. Panda et al. [7] mainly studied music emotion recognition based on audio, and they provided a very detailed account of the content of audio features. Han et al. [8] provided a comprehensive review of music emotion recognition, but they focused on the models used for music emotion recognition and provided a detailed categorization, while the description of the dataset, model evaluation metrics, and other contents were relatively simple. Although the article by Cui et al. [9] is a review of music emotion recognition based on EEG signals, they reviewed music emotion recognition from multiple perspectives, and the review is very comprehensive and rich.

Considering the lack of datasets in the field of music emotion recognition research, we present the music dataset in more detail in the Section III, and to better understand these datasets, we present some examples of commonly used datasets and include others in a table for easy comparison. In the Section IV, we introduce the commonly used emotion models, which are generally divided into two categories: category emotion models and dimensional emotion models, and we present these commonly used emotion models in the form of pictures. In the Section V, we present the common preprocessing methods for music emotion recognition downsampling, windowing, and data augmentation, and other preprocessing methods. In the Section VI of the feature extraction section, we have described audio feature extraction, and in the multi-modal feature extraction section, due to less literature related to visual and physiological features, we have only described the multi-modal feature extraction methods for the commonly used text features. In the Section VII, considering the existence of machine learning algorithms that differentiate the music emotion recognition problem into two types of problems, classification and regression, we describe the traditional machine learning algorithms used for the two types of problems, respectively. However, we focus on the narrative of deep learning models that have become very popular in the last five years, divided into unimodal and multi-modal music emotion recognition to develop. In addition, for information purposes only, we have included model evaluation methods not found in most known reviews of the field, as well as individual examples of music emotion recognition applications. The final part discusses current challenges and future research priorities in music emotion recognition. The structure of the entire review follows the general steps of music emotion recognition research, which allows for a more intuitive understanding of the steps and content of music emotion recognition research.

III. DATASETS

There are music emotion datasets is used to study the basis of music emotion recognition. As such, it is an important part of music emotion recognition. The quality, completeness, and correctness of the music emotion dataset can often affect the accuracy of the whole music emotion recognition research results data. Unlike the rich variety of speech emotion recognition datasets, music emotion recognition is in a situation of lack of datasets due to issues such as music copyright, the fact that fewer datasets are currently available, and it is more difficult and costly to create homemade datasets. Not to mention the need to find good quality, complete, and correct music audio datasets for experimental studies, the lack of datasets is one of the biggest obstacles in the field of music

emotion recognition research. Several commonly used public music emotion datasets are briefly described below.

A. EMOMUSIC DATASET

The EmoMusic dataset [10] is a dataset containing 1000 songs selected from the Free Music Archive (FMA). It collects continuous VA ratings per second, with valence indicating positive and negative emotions and arousal indicating emotional intensity, and uses dimensional representations of emotion to continuously annotate the song with VA dynamics as the song is playing. This dataset has a larger scale of emotion annotation than any existing music emotion dataset and is more suitable for music emotion recognition experiments based on the VA emotion model [11], [12], [13], [14], [15]. Excluding the redundant music dataset, about 744 music datasets can be used for music emotion recognition experiments. The dataset is divided into two parts, one for the training set of 619 songs and the other for the test set of 125 songs. The extracted 45 s excerpts are all re-encoded and had the same sampling frequency of 44100 Hz.

B. AMG1608 DATASET

The AMG1608 dataset [16] is a music emotion recognition dataset also used for VA emotion modeling, proposed and created by Chen et al. It contains 1,608 30 s music clips annotated by 665 subjects with the corresponding VA annotations, and contains annotations for 46 topics, with more than 150 music clips annotated for each topic [17]. The music emotion model uses a VA dimensional model, where each music clip is labeled with only one VA value, which can be used to analyze and study the personalization problem of emotion recognition. This dataset resource is publicly available and large, and compared to the EmoMusic dataset described above, the AMG1608 dataset also uses VA ratings, but is slightly less heavily labeled compared to the dynamic rating approach.

C. DEAM DATASET

The DEAM dataset [18] is a multi-modal music emotion recognition dataset created by Soleymani et al. They added 45 s clips of 1,744 pieces of music in 2013-2014 (consisting of music clips from 744 songs in the MediEval 2013 training set and 1,000 songs in the MediEval 2014 test set) and recorded them in stereo MP3 format at 44.1 KHz, with each clip annotated with a segment-level static VA value and a set of dynamic VA values at 0.5 s intervals. With the addition of 58 more (MediEval 2015 rating level songs) full songs averaging 4 minutes in length in 2015, today's dataset contains a total of 1,802 songs with valence and arousal annotations, making it the largest benchmark in the field of emotion recognition for continuously recognized music currently available. The emotion annotation portion of this dataset meticulously records the emotional states in the audio, allowing researchers to analyze and understand the subtle changes in emotional expression in musical compositions [19], [20], [21], [22].

D. MSD

MSD (million songs dataset) are datasets used to support commercial-scale algorithm research and to provide reference datasets for evaluation studies. This dataset has a total of 280 GB of data and mainly contains 1,000,000 contemporary popular music songs/documents, 44,745 unique artists, 7,643 Echo Nest tagged terms from, 2,321 unique music mind tags, 43,943 artists with at least one term, 2,201,916 asymmetric similarity relationships, 515,576 outdated tracks since 1922, is a large and rich dataset [23]. These data are stored in HDF5 format to efficiently handle different types of information. However, it lacks musical structure annotations and suffers from data organization and information loss problems [24]. Based on these drawbacks, Delbouys et al. [25] labeled the MSD using Deezer to obtain an MSDD based on MSD improvements. The MSDD has 18,644 tagged songs, along with their Deezer song identifier, MSD identifier, artist, and track title, and is annotated with their valence and arousal value [26].

E. CAL500 DATASET

The CAL500 dataset was created by Turnbull et al. [27] and consists of a set of 500 ''Western Pop'' songs by 500 artists, each of which contains at least three people's annotations for the song. On this basis, covering 135 music-related concepts, six semantic categories were included, including emotion categories related to music emotion recognition, with 18 emotion keywords, and each of these 18 emotions was rated on a scale of 1 to 3 (e.g., ''neutral'', ''happy'', ''unhappy''). There are a total of 1,708 annotations, and each song has a vector of binary annotations, ensuring a high degree of consistency between all topics to ensure that the tags are reliable. This dataset basically satisfies the fine-grained and discriminative aspects required for music emotion recognition. In addition, based on the CAL500 dataset, Wang et al. [28] published the well-known CAL500 enriched version, CAL500exp. In contrast to the track-level annotations of CAL500, CAL500exp tags are annotated at the segment level, meaning that each song contains several segments separated from itself. And each segment was labeled as dependent data from 18 sentiment labels, resulting in a more focused CAL500exp data with 3223 items [29].

F. PMEMO DATASET

The PMEmo dataset contains emotion annotations for 794 songs and electrodermal activity (EDA) signals synchronized with song auditions to facilitate multi-modal emotion recognition. It was created by Wang et al. [30] who recruited 457 subjects for VA annotation. The 794 songs of varying length were taken from the Billboard Hot 100, iTunes Top 100 Songs (U.S.), and the U.K. Top 40 Singles Chart, all recorded in 44.1KHz stereo MP3 format. In addition to the aforementioned songs with VA annotations and EDA signals, the PMEmo dataset contains song metadata (song title, artist, timestamps of the beginning, and end parts of the chorus),

TABLE 2. By investigating known music emotion datasets.

Datasets	Contents	modal	Language	Publisher	Year
EmoMusic [10]-[15]	1000 songs, VA annotations	unimodal	multi-language	Soleymani *et al.*	2013
AMG1608 [16][17]	1608 songs, VA annotations	unimodal	——	Chen *et al.*	2015
DEAM [18]-[22]	1802 songs, VA annotations	unimodal	multi-language	Soleymani *et al.*	2013-2015
MSD [23]-[26]	Derived features of 1 million songs, profiling, song metadata	unimodal	——	Bertin-Mahieux *et al.*	2011
CAL500 [27]-[29]	500 songs, audio representation, emotional annotations	unimodal	multi-language	Turnbull *et al.*	2008
PMEmo [21][30]-[32]	794 songs, emotional annotations, song metadata, pre-calculated audio features, synchronized EDA physiological signals, lyrics, the song comments, extended physiological feature space	multi-modal	——	Zhang *et al.*	2018-2019
DEAP [33]	40 music videos, synchronized computerized electroencephalogram (EEG), arousal, value, preference, dominance, and familiarity ratings, 22 participant frontal facial expression videos	multi-modal	——	Koelstra *et al.*	2012
ASDB [38]	80 samples of Assamese music	unimodal	Assamese	Dutta *et al.*	2021
Turkish Emotional Music Database [39]	124 music clips of 30 s duration	unimodal	Turkish	Hizlisoy *et al.*	2021
VioMusic [40]	264 violin solos, 1926 music fragments, VA model sheet music, feature data	unimodal	——	Ma *et al.*	2024
EMOPIA [41]	1087 music clips, MIDI files	multi-modal	——	Hung *et al.*	2021
GTZAN [42]	1000 songs	unimodal	English	Tzanetakis *et al.*	2002
TROMPA-music emotion recognition [43]	4721 music annotations, 1161 music clip information	unimodal	multi-language	Gómez-Cañón *et al.*	2022
4Q audio emotion [44]-[46]	900 music clips, VA annotations	unimodal	——	Panda *et al.*	2018
Bi-modal emotion [45]-[47]	133 music clips, lyrics, VA annotations	multi-modal	——	Malheiro *et al.*	2016
PSIC3839 [48][49]	3839 songs, VA annotations, 2372 sets of lyrics	unimodal	Chinese	Xu *et al.*	2022
Hindi songs dataset [50]	1000 songs	unimodal	Hindi	Chaudhary *et al.*	2021
RAVDESS [46]	848 music clips	unimodal	——	Livingstone *et al.*	2018
EmoMV [51]	EmoMV-A: 4916 music video clips divided into training, validation and test sets, EmoMV-B: 616 music video clips divided into training and validation sets, EmoMV-C: 456 music video clips divided into training and validation sets; matching music video pairs and mismatching music video pairs; 5 emotional labels	multi-modal	English	Thao *et al.*	2023
MERP [52]	54 full songs, VA dynamic scoring, personal information such as annotator's musical preferences and musical backgrounds	unimodal	English	Koh *et al.*	2022

pre-computed audio features for the music emotion recognition task, and even lyrics and user comments. In the PMEmo dataset, for the static emotion task, 6373-dimensional song level features are provided for the audio features, while for the dynamic recognition task, Wang et al. extracted only 260-dimensional segment-level kernel features, and all features were extracted by the OpenSMILE open-source toolkit.

G. DEAP DATASET

The DEAP dataset [33] is a multi-modal dataset based on the analysis of human emotional states. The dataset contains physiological cues from 32 participants and frontal face videos from 22 participants. Each participant watched 40 music videos based on video arousal, valence, dominance, and perception, and rated emotional responses based on preference and familiarity. The dataset is selected as a semi-automatic stimulus selection method based on affective labeling and is open to the research community, which is of great significance to affective research efforts in many fields.

H. OTHER DATASETS

In addition to the widely used datasets mentioned above, for the music emotion recognition task, each researcher uses different algorithms, goals, or frameworks, and has certain

requirements for the suitability of the dataset, so some people choose to create their own datasets or other more niche datasets for experimentation. For example, Panda et al. [34] organized their dataset based on the AllMusic database in a manner similar to the MIREX Mood Classification Task testbed; Dufour et al.'s [35] dataset was selected from a personal music library of nearly 600 songs and included a variety of music genres in different languages; Baniya et al. [36] experimentally selected the Jyu dataset. There are also datasets for music emotion recognition in different languages, Korean Naver Music dataset [37], Assamese ASDB dataset [38], Turkish dataset [39], as well as datasets for pure music emotion recognition, violin VioMusic dataset [40], piano EMOPIA dataset [41], and other datasets. Without going into all the details, we present a summary in the form of Table 2.

IV. MUSIC EMOTION MODELS

In order to successfully implement a music emotion recognition system, a music emotion model needs to be used in the music emotion recognition process. However, there is no consensus on the correct choice of emotion model, which is closely related to the type of underlying factual data in the experiment. As far as the current general scientific research results in the field of music emotion recognition are concerned, the types of music emotion models can be categorized into category emotion models (discrete emotion models) and dimensional emotion models (continuous emotion models).

A. DEFINITION OF MUSIC EMOTION

In the field of music emotion recognition, researchers have broadly categorized the definition of emotions in music into ''Expression Theory'' and ''Arousal Theory''. The ''Expression Theory'' holds that the emotions contained in music are conveyed by the performers through the music, which is their emotional expression, while the ''Arousal Theory'' holds that the emotions contained in music are the emotions that the listener feels by listening to the music. This shows that ''Expression Theory'' tends to convey emotions from the perspective of the music creator, while ''Arousal Theory'' tends to feel emotions from the perspective of the listener. The results of many related studies show that different people's choices of emotions expressed by the same music are mostly the same, and this method of judging musical emotions based on people's perceptions has been proven to be reliable and effective [120]. However, the expression of musical emotion is a complex issue, unlike the general public in the past that low music makes people feel sad and melancholy, soothing music makes people feel calm, exciting music gives people inspiration and exhilaration, in many music, happy music tone will be mixed with sad, sad music tone will be mixed with cheerful and other cases. For example, the upbeat melody of ''Viva La Vida'' tells the sad story of the rise and fall of a king, while the sadness of the intro to ''Hey Jude'' becomes progressively more upbeat as the song progresses. Different listeners tend to hear different emotions in the same song, so there is a difference between the emotions in the ''Expression Theory'' and the ''Arousal Theory''. In music emotion recognition research, the labeling of music emotion is based on ''Arousal Theory'', which selects the listener's customized emotion as the emotion label for the study, and inevitably there will be individual differences. Therefore, adopting a relatively objective ''Expression Theory'' and selecting the emotion label specified by composers and performers is more consistent with the objectivity of scientific research in terms of application.

B. CATEGORY EMOTION MODEL

For the category emotion model, the focus is on distinguishing the characteristics of emotions that people use in their daily lives to define the emotions they observe, and thus the labeling scheme based on emotion categories is intuitive. Henver's emotion ring [53] is one of the earliest and most influential music emotion models among discrete emotion models. The Geneva Music Scale (GEMS) [54] is the first model to measure the emotions of music listeners specifically from their point of view, and it contains 45 emotion labels grouped into nine categories, i.e., wonder, transcendence, tenderness, nostalgia, peacefulness, power, joyful activation, tension, and sadness. In Ekman's [55] article, emotions are summarized in general terms as anger, disgust, fear, happiness, sadness, and surprise. In 1980, the psychologist Plutchik also first published his model of the emotional wheel Fig. 2, in which he proposed four pairs of two-level emotions: joy and sadness, anger and fear, trust and disgust, and surprise and anticipation, which is also a category model. Roseman et al. [56] assessed emotions through evaluative factors, giving 17 basic emotions. Most existing music emotion recognition systems focus on these basic emotion categories. Therefore, in order to better fit the experiment and measure the data, many researchers reduce the widely used basic emotions or add entirely new combinations of discrete emotions to conduct the experiment [57], [58], [59], [60], [61]. Panda et al. [34] used five discrete clusters of emotion defined according to MIREX in their emotion recognition task Fig. 3 for the experiment. Feng et al. [57] developed a category model consisting of four emotions that are discrete emotions, namely happy, sad, angry, and fearful basic emotions; Lu et al. [58] divided changing emotions into separate parts, each of which contained stable emotions; Er and Aydilek [59] selected a dataset with emotion labels and used Softmax to categorize into four emotion categories: anger, sadness, happiness, and relaxed, and Zhang et al. [60] similarly categorized into the above four categories; Jeon et al. [37] distinguished between positive and negative labels according to a predefined lexicon of emotion words, i.e., positive emotions such as ''happy'' and ''cheerful'' and negative emotions such as ''sad'' and ''lonely'' as two separate broad categories; Jia [61] used Softmax for classification to output four discrete emotions, namely happy, calm, healing, and sad. In the following, we will discuss the most classical Hevner emotion model.

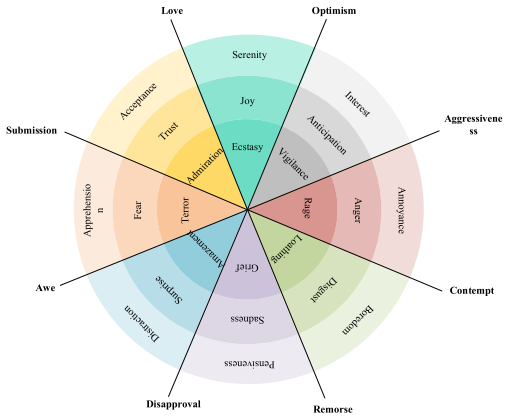


FIGURE 2. Plutchik emotion wheel.

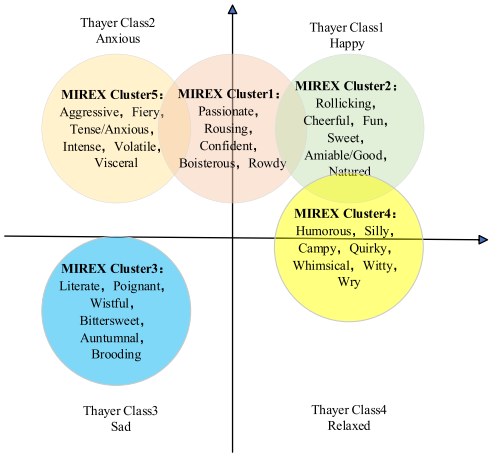


FIGURE 3. Panda et al. cluster of emotions.

Hevner's model is a psychological model of music emotion that is often used or referenced in the field of music emotion recognition. It belongs to the category emotion model, first proposed by Hevner in 1936 Fig. 4. The model attributes the 67 emotion adjectives found by Hevner to eight clusters, namely solemnly, sad, chant, desire, lyrical, bouncing, happy, and passionate. Each cluster is subdivided into several more detailed and broader adjectives below it, and the adjectives between clusters are similar. And neighboring clusters in the Hevner emotion ring are cumulatively varying in significance, with two opposite clusters having opposite significance. Hevner explained the model as a more or less continuous scale that takes into account minor disagreements between annotators and the influence of pre-existing emotional or physiological conditions that may affect the annotators' perceptions, which is one of the strengths of the model [35]. Farnsworth attempted to improve consistency within and between clusters by changing some of the adjectives in Hevner's emotion model. He then reorganized some of these clusters and added some new adjectives in 1954 [62], and revised the Hevner model again in 1958, but these revisions ignored the Hevner emotion ring [63]. The Hevner model has not been updated since Farnsworth's last revision in 1958. In 2003, Schubert [64] asked 133 musically experienced people about the applicability of 91 adjectives describing music. These adjectives were taken from the original 67 words of Hevner's model, as well as additional words from Russell's model and Whissell's emotion lexicon, which were added and subtracted from Hevner and Farnsworth's results to arrive at a discrete emotion model of nine emotion clusters and 46 adjectives.

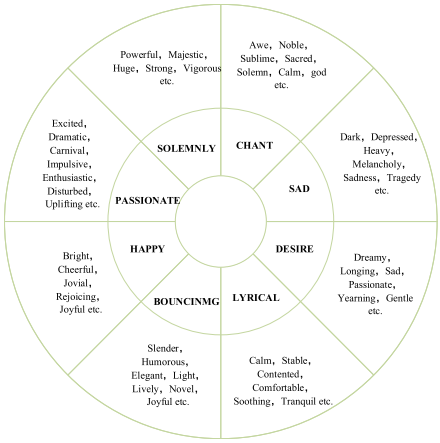


FIGURE 4. Henver emotion ring.

C. DIMENSIONAL EMOTION MODEL

The main drawback of the category emotion model is that the number of primary emotion categories is too small compared to the richness of musical emotions that humans can perceive. On the other hand, using finer granularity does not necessarily solve the problem because the language used to describe emotions is inherently ambiguous and varies from person to person. Because the above discrete emotion categories do not define some of the complex emotional states observed in everyday communication, dimensional emotion modeling provides an alternative to music emotion recognition. In contrast to the category emotion model, the dimensional emotion model focuses primarily on recognizing emotions based on their location on a small number of emotion dimensions with named axes that correspond to internal human representations of emotions, thus, the dimensional model represents human emotions as coordinate points in a multi-dimensional continuous space. In general, the higher the dimensionality, the more detailed the emotion expression, but the fact is that the higher the dimensionality, the more complex the emotion analysis becomes, and therefore the dominant dimensional models are two- or three-dimensional emotion models.

1) RUSSELL EMOTION MODEL

Russell's emotion model [65] is one of the most typical two-dimensional emotion models, namely the widely used VA model [66], [67], [68], [69], [70], [71], [72], [73], [74], [75]. His proposed dimensional model abstracts emotions into a continuous dimensional space consisting of two axes,

Valence and Arousal Fig. 5(a). Where the V-axis is used to represent the concepts of negative and positive emotions, and the A-axis is a measure of the level of stimulation of the emotion by the music. In this model, emotions are placed away from the starting point because that is where valence and arousal are higher, and therefore emotions are clearer. Panda et al. [69] pointed out that since the Valence axis and Arousal axis are separate and independent, the Russell emotion model can be considered discrete, i.e., discrete and dimensional emotion representations can be interconverted to some extent. Therefore, some researchers [26], [43], [76], [77], [78] make improvements based on Russell's emotion model to avoid overly complex or unvalidated categorization of emotions. They grouped all emotions into a quadrant of a Russell emotion model and derived four categories of emotions corresponding to the four quadrants, such as happy, angry, sad, and relaxed; angry, surprised, fearful, and sad; high-positive, high-negative, low-positive, and low-negative emotions; and happy, tense, sad, and calm.

2) THAYER EMOTION MODEL

Thayer's model is a two-dimensional, dimensional model of emotion formalized by psychologist Thayer in 1989 [79]. The Thayer model is based on an improvement of the Russell model. Unlike Russell's model, where the two axes are defined as Valence and Arousal, respectively, Thayer's model of emotion puts a 45-degree spin on Russell's. Both axes of the Thayer model represent arousal, one axis representing energy arousal and the other representing pressure arousal Fig. 5(b). Because in Thayer's view, the two potential dimensions should be two separate dimensions of arousal, and the valence should be a combination of energy arousal and pressure arousal. In Thayer's model, energy serves as the vertical axis, evaluating indicators from calm to vitality, and stress is the horizontal axis, evaluating indicators from happiness to worry. The two-dimensional plane of emotion is divided by these two horizontal and vertical axes into four quadrants corresponding to the four extremes of emotion-anxiety, exuberance, contentment, and depression. The advantage of the existence of this model is that Thayer suggested from a psychological point of view, described in terms of dimensional thinking, that both energy and stress factors are more compatible in terms of their correspondence with acoustic features, and that they can be well connected to auditory features. Therefore, the Thayer emotion model is often used in current emotion studies of musical works in MP3 and WAV formats.

3) TWC EMOTION MODEL

However, Thayer's model falls short when faced with a large and complex vocabulary of emotion descriptions. And as for dimensional modeling, it has been controversial in the past because of the lack of distinction between immediately adjacent emotions (e.g., anger and fear) on the valence and arousal dimensions. Therefore, Tellegen et al. [80] made a modified extension of the Thayer dimensional model and proposed the TWC (Tellegen-Watson-Clark) model. The model integrates existing emotion models and stratifies emotions, with the valence dimension (pleasant-unpleasant) at the top, two separate horizontal and vertical axes in the middle - the negative emotion axis (low negativity to high negativity) and the positive emotion axis (low positivity to high positivity), and specific emotions at the bottom, including joy, sadness, guilt, fear, and other emotions Fig. 5(c). The TWC model not only retains the natural and smooth emotional transitions of the Thayer model, but also greatly enriches the words used to describe musical emotions.

4) PAD EMOTION MODEL

The PAD model is a three-dimensional measurement model first proposed by Mehrabian and Russell [81], [82]. The three letters of PAD correspond to the three axes of the three-dimensional model Fig. 5(d), which are Pleasure (representing the positive and negative features of an individual emotional state), Arousal (representing the level of an individual neurophysiological arousal), and Dominance (representing a sense of control and influence over one's surroundings and others). According to these three dimensions, emotions can be categorized into eight categories: happy, angry, disgusted, relaxed, fearful, sad, surprised, and satisfied. The PAD model is based on an improvement of the VA model and compensates for the shortcomings of the VA model by being able to differentiate between neighboring emotions such as anger and fear. The PAD model can describe an emotion intensity state based on specific coordinates in space and can provide a continuous, nuanced description of emotion words. In contrast, Thayer's model has only two dimensions and lacks emotional richness, while Hevner's emotion ring model categorizes and describes emotions in too much detail, in addition to the 8-dimensional emotion ring, each category is subdivided into several subcategories, which is too detailed a categorization. Simple music with a single emotion does not require complex semantic modeling. Although Bakker et al. [83] pointed out that environmental psychologists still question the interpretation of the pleasure, arousal, and dominance dimensions of the PAD emotion model, as well as the underlying mechanisms, the model proposed by Mehrabian et al. continues to be widely used in research applied to psychology, marketing, and product satisfaction.

V. DATA PREPROCESSING

A. DOWNSAMPLING AND FRAMING

The purpose of down-sampling is to reduce the amount of data and computational complexity, while signal framing is used to divide the audio signal into smaller segments for feature extraction and emotion analysis. If the audio content of a song is a few minutes long, the emotion it contains will fluctuate briefly, causing the results of the experiment to become less accurate. To avoid this and to improve the

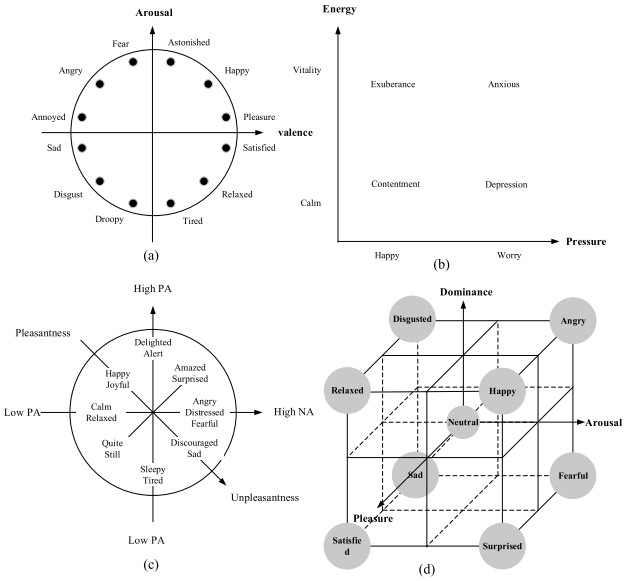


FIGURE 5. Commonly used dimensional emotion models.

accuracy of music emotion recognition, researchers typically divide a long piece of music into several music segments and recognize the emotions contained in the music in segments. There is often a consensus on the length of segments used to segment music. For example, the typical segment length for popular music is 25-30 s [6], while the optimal segment length for classical music is 8-16 s [84]. Nevertheless, the choice of the length of the musical segments has to be carefully selected on the basis of the specifics of the experiment at the time, as it needs to take into account the validity of the emotion response and the homogeneity of the learning of the features of the segments, as well as consider the adaptability of the model, among other factors. For this reason, He et al. [31] compared the experimental effects of four types of music test clips with durations of 1 s, 3 s, 5 s, and 10 s to select the most appropriate music clip length for the experiment. Lucia-Mulas et al. [85] selected 94 clips from their dataset that had an original sampling rate of 44.1 KHz, were downsampled to 16 KHz, and were segmented into 976 samples of 2 s from the original 94 clips with an average duration of 16 s. They thought that 2 s would be enough to produce immediate music emotion, which would enable the experimental results to be produced more efficiently, and the final results proved this. Han et al. [86] then cut music clips strictly to a duration of 30 s, and samples of less than 30 s were supplemented by copying the original audio using an audio editing tool. Since deep neural network models require a large amount of data, Gupta [45] solved the problem of scarcity of music data to some extent by splitting each 30 s sample sampled at 22050 Hz mono into three 10 s segments, removing invalid data less than 8 s to increase the data used to train the model, and providing higher efficiency for the model. Chen and Li [87] to be more effective in extracting musical audio features with excessive feature size and complex synthesis. After preprocessing operations such as signal framing of audio samples, fine-grained segmentation and vocal separation, they constructed four experimental datasets, namely 30 s original, 15 s original, 15 s pure background, 15 s pure vocal. The classification performance of these four datasets was validated by low-level descriptors with support vector machine classifiers, and it was finally found that 15 s audio clips had higher fine-grained accuracy than 30 s, and the average classification accuracy of pure background audio was higher than that of pure vocal audio, so 15 s pure background audio clips were chosen as the dataset for the experiments.

B. WINDOW FUNCTIONS

Windowing is an important step in the processing of music audio signals, the main purpose of which is to reduce spectral leakage and fence effects. Windowing a music signal is a direct application of the window function (1) to deal with the large peaks that often occur in music signals, and the dot product of the music time series is used to smooth the signal.

$$W_M(n) = \begin{cases} W(n) & 0 \le n \le M-1 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

In (1), M in the window function refers to the window length and the range of values of n is $0 \le n \le M-1$. Window functions that are commonly used are Hanning window, Hemming window, rectangular window, and triangular window. Xia and Xu [88] used the Hanning window function (2) in their experiments to realize the short-time frame window processing of the music signal. Han et al. [86] also used a Hanning window for data preprocessing.

$$W(n) = 0.5 - 0.5\cos(\frac{2\pi n}{M-1}) \tag{2}$$

They stacked the information of each frame on the time axis by moving the window function on the time axis, and then further processed the signal after adding the window, such as fast Fourier transform, MEL cepstrum coefficient calculation, and other processes, to finally get the audio spectrograms needed for the experiment. Jia [89] used the Blackman-Harris window function (3) to divide the original signal sampling rate of 44.1 KHz into 2048 sampling frames.

$$W(n) = 0.42 - 0.5\cos(\frac{2\pi n}{M-1}) + 0.08\cos(\frac{4\pi n}{M-1}) \tag{3}$$

C. DATA AUGMENTATION

Data augmentation is also often used in data preprocessing. In music emotion recognition tasks, a large number of datasets are often required to train the proposed model in order to minimize the error between the training data and the test data. However, as we mentioned earlier, there is an occlusion in the circulation of music datasets, there are fewer music datasets available for training, and most of the music datasets are limited in content, and the creation of additional data training samples through data augmentation can compensate for the problem of the limited data size to some extent. Er and Aydilek [59] varied the audio playback time for

audio samples in the dataset by varying its multiplier speed to 0.81, 0.93, 1.07, 1.23. And they did this by moving the start of each audio sample to a position five seconds later while keeping the total duration constant, resulting in a new dataset six times the size of the original one. In Chang et al.'s [21] music emotion recognition experiments, two data augmentation methods were used to preprocess the dataset. One is to multiply the amplitude of the music signal by a randomly chosen scale in the interval [0.9, 1.1], and the other is to add zero-mean Gaussian noise ($\sigma = 0.02$) to the signal to achieve enhanced variability in the training data. Tong [90] used three processing methods for data augmentation. The first was in the audio preprocessing part of all the audio to take a Gaussian noise for audio noise augmentation, so as to offset to a certain extent the song recording and propagation process accompanied by a variety of noise; the second is the audio was cropped to 30s segments; the third was a random mix of different audio, the two segments of the audio for the mixing of the generation of mixed samples could be directly used to calculate the model KL loss.

D. FEATURE DETECTION AND SELECTION

Feature selection is a method of reducing the size of extracted features by selecting the most salient ones. Hizlisoy et al. [39] used correlation feature selection (CFS) for feature selection, which is an algorithm that evaluates a subset of features using an objective function to find out the irrelevant and relevant subsets. CFS uses symmetric information gain to calculate tff and tcf, tff is the average feature-to-feature correlation, tcf is the average class-to-feature correlation, and n is the number of features in the F-feature subspace, and its algorithm is as follows:

$$\text{Merit}_F = \frac{nt_{cf}}{\sqrt{n + n(n-1)t_{ff}}} \tag{4}$$

Jandaghian et al. [91] proposed a fuzzy system to discover the most effective features for each emotion among the music segments. The system got the value of each component in the music feature vector and selected the feature with the highest value. Krols et al. [92] tested the importance of audio features in multi-variate linear regression and obtained five audio features with significant effects: danceability, energy, instrumentalness, valence, and mode. For the selection of lyric features, they used the comparison of VA scores of any combination of computed emotion features, TF-IDF features, and Aenew features to select the combination of features that would best improve the experiment. In addition, they applied to recursive feature elimination (RFE) for multi-modal feature selection. Juthi et al. [77] used a random forest feature ranking algorithm, which is also a method used for feature selection. By ranking the eight music features evaluated, energy, roll-off, tempo, brightness, pitch, key, mode, and MFCC, they found that roll-off was the only feature that scored equally high except for the arousal feature, energy, which showed a high priority among the all features. The ReliefF feature selection algorithm used by Panda et al. [44] is similar in principle to the random forest feature ranking method, but ReliefF focuses on the ability of the features to discriminate between nearby samples, while random forest focuses on the degree to which the features contribute to the overall model prediction performance. The shrinkage method is also one of the feature selection methods, comparing the two feature selection methods, filter and wrapper, which are especially excellent in selecting the arousal dimension features. Two common shrinkage methods currently used are ridge regression and LASSO regression. Dong et al. [22] considered that ridge regression could not produce sparse models, making it difficult to detect emotionally salient features, so LASSO regression was chosen to set the coefficients of some neurons to zero for the selection of salient features to achieve the purpose of removing redundant variables and noise.

E. DIMENSION REDUCTION AND NORMALIZATION

The main purpose of feature dimension reduction is to reduce the computational complexity of the model, improve model performance, and remove redundant information. However, sometimes it also filters out the useful information in the audio features. In order to maximize the retention of useful information in the audio features and effectively retain the main features of the audio, Han et al. [86] first performed dimensional reduction of the features into the learning network to reduce the learning burden of the learning network, and then expanded the features to restore the original dimensions to rebuild the output of the features, which formed the input feature compression-expansion-compression structure. Xia et al. [88] used principal component analysis (PCA) for feature dimension reduction, which is a statistical method for processing, compressing, and extracting information from samples based on variable covariance on the feature matrix. Its core idea is to project the data along the largest direction to make the data is easy to distinguish. It can be as much as possible to retain the original reflection of the information at the same time and effectively reduce the number of features containing noise or redundancy, which is a commonly used dimension reduction method. Batch normalization is a technique used to improve the performance of deep learning networks by normalizing the interlayer outputs of a neural network based on supervised learning. One of them, Batch Norm2d, has been used mainly as a convolutional layer of neural networks, which is particularly suitable for processing 2D data. Ma and Zhou [40] applied Batch Norm2d to normalize the 2D convolutional layer of the input emotion feature matrix in violin music emotion recognition to prevent model overfitting and improve model performance and stability.

VI. FEATURE EXTRACTION

A. AUDIO FEATURE EXTRACTION

1) COMMON AUDIO FEATURES

Audio information mainly includes time domain features, frequency domain features, and other features. Time domain

features are notable in that they do not require any form of transformation of the original audio signal, but are processed on the sampled values of the signal itself, which includes zero crossing rate (ZCR) features, energy features (including root mean square energy), time domain centroid, and other features; frequency domain features are usually closely related to timbre, and can be categorized in general as Fourier transform (FFT)-based features and short-time Fourier transform (STFT)-based features, and also in general as spectral envelope-related features, spectral structure-related features, statistics-type features, and coefficients features. In addition to the time and frequency domain features mentioned above, Mel frequency cepstrum coefficients (MFCC), linear predictive coefficients (LPC), and perceptual linear predictive coefficients (PLP) among other features, are also used in music emotion recognition. However, in our survey of the literature in recent years, the MFCC feature has been the most widely used. Because MFCC contains more audio information among the low-level features of audio, it simulates the auditory perception of the human ear and is able to extract useful music features that are more important to human perception. And since there is less literature related to the application of LPC and PLP in the feature extraction part, we will focus on the description of MFCC related content next.

2) MEL SPECTROGRAM FEATURE AND MFCC FEATURE

The successful application of neural networks in the field of image recognition in recent years has led researchers to explore the feasibility of applying neural networks for image recognition to music emotion recognition. The model proposed by Li et al. [93] using convolutional neural networks (CNN) for feature extraction and classification of Mel spectrograms containing musical features showed good feature extraction performance, confirming this. In contrast to the spectrum, which can only describe the distribution of sound at each frequency at one point in time, the spectrogram can describe the situation over a period of time, and it is the continuous generation of the spectrum that can more intuitively show the changes of the audio signal in time and frequency. One after another, researchers began to adopt the model of transforming the Mel spectrogram of audio to convert one-dimensional audio into two-dimensional images as input to the neural network. The Mel spectrogram is obtained by pre-emphasizing the signal and then performing a short-time Fourier transform followed by Mel filtering. When the Mel spectrogram is further processed by taking logarithms, separating the signals, and performing a discrete cosine transform, the familiar Mel frequency cepstrum coefficients (MFCC) are obtained. Dutta and Chanda [38] analyzed and compared the effect of MFCC and chroma features on the emotion performance of music in their proposed music emotion recognition model for Assamese music. They found that the accuracy of MFCC was much higher than that of chroma features, by 31.25 percent. It can be seen that the features obtained by MFCC have a very positive impact on the experimental results, greatly improving the ability to classify musical emotions. It is worth mentioning that most of the features of the audio sample in the MFCC extraction technique are concentrated in its first few coefficients, as also mentioned in Dutta et al. This is because the more advanced coefficients contain more information about the fundamental frequency and resonance peaks, which is also a characteristic of the ''energy concentration'' of the discrete cosine transform. The feature representation section of Zhang et al. [94] mentioned the extraction of filter bank outputs from the original audio signal to represent the original features of the audio signal, which were input to the proposed model. This filter bank output is intermediate between the Mel spectrogram and the MFCC. Comparing the Mel spectrogram and MFCC, the output of the filter bank eliminates some of the redundant information and retains more details, which reduces the complexity of the experimental feature extraction and improves the classification performance of music emotion. Considering that the lack of information will make the deep neural network prone to overfitting, and the defects that the spectrogram can only represent the information in the visual domain while ignoring the statistical data, Zhang et al. [14] added a frequency embedding module to supplement the numerical data details in the proposed music emotion recognition network structure. They first transformed the obtained spectrogram (5) to obtain the Mel frequency cepstrum, then obtained the Mel cepstrum coefficients by discrete cosine transform (DCT) (6), and finally obtained the transformed spectrogram by frequency embedding calculation (7) and input it into the feature extraction model.

$$\mathrm{Mel}(m) = \sum_{k=t(m-1)}^{t(m+1)} \mathrm{Fil}_m(k)|F(k)|^2 \tag{5}$$

$$\mathrm{MFCCs}(m) = \mathrm{DCT}(\log(\mathrm{Mel}(m))) \tag{6}$$

$$\mathrm{EmbdFreq} = \mathrm{FFT} \oplus \mathrm{Mel} \oplus \mathrm{MFCCs} \tag{7}$$

In (5), (6), and(7), Mel is the Mel frequency cepstrum, m is their first filter, $F(k)$ is the energy spectrum computed by FFT, $t(m)$ is the transform equation between Mel frequency and Hertz frequency, log denotes logarithmic operation, and symbols $\oplus$ denote connection operations in the channel. To improve the capability of the model for feature extraction, Jia [89] applied the residual phase (RP) to the feature extraction part, which can extract audio specific information that complements the MFCC features. Thus, he performed a weighted combination of MFCC features and RP features to obtain the final low-level feature output Fig. 6.

In addition to using the MEL spectrogram, Lucia-Mulas et al. [85] had also used STFT (short-time Fourier transform) and CQT (constant-Q-transform) spectrograms in conjunction with convolutional neural networks in several experiments. Er and Aydilek [59] used MIRToolbox for chroma extraction of spectrograms for music emotion recognition in their experiments. Niu [95] extracted multiple spectral features in music to form a sequence that captures

feature information from the music essence for the feature extraction part of music emotion recognition. In particular, he added a spectral feature for rhythmic features in music, extracted using the beat histogram, which passed the time domain feature through the results of several filters to extract a 16-dimensional feature. Louro et al. [96] argued that for training machine learning models for emotion classification tasks, adopting Mel spectrograms may not be the optimal solution. Considering that embedding is widely used in natural language processing (NLP), Koh and Dubnov [46] divided the music emotion classification task into two steps Fig. 7. In the first step, a piece of music was taken as input and the deep audio embedding, which was indicative of the acoustic features of the music, was obtained using L3-Net or VGGish deep audio embedding model Fig. 8. The second step was to select traditional machine learning and deep learning models for classification after obtaining deep audio embeddings, respectively. The final results showed that L3-Net performs better than the baseline MFCC features across multiple datasets.

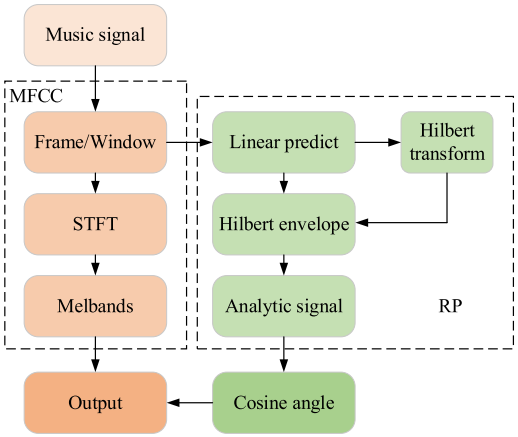


FIGURE 6. Combination of MFCC feature and RP feature.

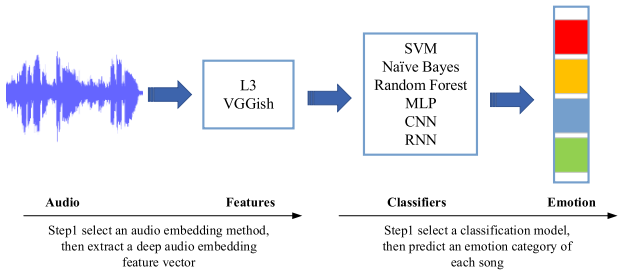


FIGURE 7. The model proposed by Koh et al.

3) AUDIO FEATURE EXTRACTION TOOL

The aforementioned L3-Net is generated based on an open-source feature extraction tool, OpenL3 [45]. There are many similar feature extraction tools, such as Librosa, Yaafe, Aubio, Vamp-plugins, and Madmom, and these libraries basically support the Python language. However, the Matlab feature extraction toolbox MIRtoolbox and the Java implementation of the feature extraction project jAudio used in Hizlisoy et al.'s [39] work on standard audio feature extraction do not directly support the Python language. In order to reduce the tediousness of the experimental process, for feature extraction, researchers tend to choose appropriate tools to perform feature extraction directly. Grekow [42], [72] and Dufour and Tzanetakis [35] among the feature extraction tools have chosen Marsyas feature extraction software, written by George Tzanetakis, which extracts 31 features such as zero-crossings, spectrum centroid, MFCC, and chroma feature. Also used in Grekow's research is another feature extraction tool, Essentia, an open-source function library that contains a large number of functions for extracting spectral features, rhythmic features, pitch features, and other features. Xie et al. [12] introduced a new approach for feature extraction. Based on the ability of sinusoidal transform coding (STC) to accurately represent the spectrum of high frequency content and preserve the high frequency components, they extracted the features representing the spectral shapes and envelopes of the music signals using STC, and used the high-level features extracted by OpenSMILE as the baseline features for the experiments as one of the comparisons. They also proposed to combine the STC extracted features with the baseline features to input them into the model to perform the music emotion classification task. In the end, the combination of baseline features and STC-based features produced the best results in terms of both RMS error calculation results and Pearson correlation coefficient calculation results. The aforementioned OpenSMILE is also one of the open-source toolkits commonly used for automatic feature extraction from audio signals and classification of speech and music signals.

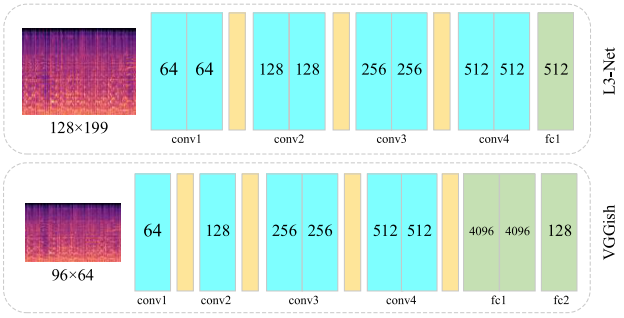


FIGURE 8. Schematic of L3-Net and VGGish deep audio embedding network structure.

B. TEXT FEATURE EXTRACTION

Regarding the music emotion recognition task, the features involved are music audio features, lyric text features, video visual features and human physiological features, but there are fewer extraction methods related to visual features and physiological features, so we mainly describe the extraction of text features. Text features include lyrics, song title, creator, album name, and other text, but mainly focus on text features of lyrics. Bradley and Lang [97] found experimentally that textual features of lyrics tend to outperform

audio features in emotion classification tasks. Krols et al. [92] created three features about the lyrics, the emotion feature, the TF-IDF feature, and the Anew feature, to predict the valence and arousal of the song as a way to represent the lyrical information. Among them is TF-IDF, which was also used by Jia [61] in his experiments. The extraction of word frequency features using text frequency-inverse document frequency (TF-IDF) is a commonly used method for text analysis. Term frequency (TF) and inverse document frequency (IDF) are calculated in a probabilistic statistical manner and their product is used to evaluate the importance of a word in a document, the higher the value, the more important it is. However, the method is flawed and ignores textual integrity, which can lead to the loss of fine-grained emotional semantic content. Su et al. [98] proposed Word2vec efficient word embedding model, which mainly includes continued bag of words (CBOW) and skip-gram two model structure. The former is used to represent a priori probabilities and the latter is used to represent posterior probabilities. In the MMD-MII multi-modal music emotion classification model proposed by Wang et al. [20], VGGish was used for audio feature extraction, and ALBERT for text feature extraction. ALBERT is a Transformer architecture based on BERT that has fewer parameters than BERT, is more lightweight than BERT, and still performs very well at various feature extraction tasks despite the content shrinkage. In Wang et al.'s model, they also took advantage of the ability of ALBERT to recognize complex relationships between words and contextual cues, and integrated the contextual information and emotional nuances of the lyric text to enrich the model's understanding of the emotional depth and complexity of the lyric content. Unlike the randomness of BERT, the XLNet used by Sams and Zahra [99] calculated all possible orders of each token to make predictions about words in sequential order, so this approach allows more textual information to be available for the overall model. The language indexing and word count (LIWC) package, published by Pennebaker [100] in 2001, can be used to analyze lyric texts in more than 70 language dimensions. In Xu et al.'s [48] study on the relationship between lyric features and perceived emotion of Chinese songs, they performed the preprocessing operations of removing redundant information such as lyricist and composer, and segmenting the lyrics text by using a Chinese word segmentation tool before extracting lyric features, and adopted a simplified Chinese version of SC-LIWC based on the improved LIWC. It extends the text features to more than 100 dimensions, from which it extracts a total of 98 lyric features, including the total number of words, the proportion of positive emotion words, the proportion of profanity, and other lyric features.

VII. MUSIC EMOTION RECOGNITION MODELS

A. TRADITIONAL MACHINE LEARNING MODELS

The music emotion recognition problem can be studied as either a classification problem or a regression problem. The problem of classification is generally studied using a combination of category models of emotion, which categorize the of emotion in music into several fixed categories of emotion. Regarding the regression problem is generally used to combine the dimensional model of emotion for research, the most common is to find the corresponding emotional categories by calculating the VA values (Valence and Arousal) mapped to the two-dimensional coordinates among the music emotions. Of course, there is also calculation of PAD (Pleasure, Arousal, and Dominance) values mapped to the three-dimensional coordinates to assess emotion. These emotion models have been described in some detail above. In traditional machine learning models, to deal with classification problems or regression problems, people often use the corresponding classification algorithms or regression algorithms.

1) MODELS FOR SOLVING CLASSIFICATION PROBLEMS

Panda et al. [34] used support vector machines (SVM), k-nearest neighbors (K-NN), C4.5, and naive Bayes algorithms for performance comparison in music emotion recognition study. The parameter-optimized SVM achieved the best results in terms of emotion recognition accuracies of 46.3%, 59.1%, and 64.0% for both standard and melodic audio features and the combination of the two as model inputs, respectively. It can be seen that the performance of SVM is superior among many traditional classifiers and is therefore widely recognized. Xu et al. [101] also chose binary SVM as a classifier. They used LibSVM software to perform the music emotion classification task by training it in a one-to-many fashion, and also optimized the SVM parameters by incorporating a radial basis function (RBF) kernel for multiple cross-validations of the training data. The original audio source separates the accompaniment and the song and then fuses them at a later stage, which can improve the recognition accuracy of the whole classifier to 0.532. Chin et al. [71] proposed a two-layer SVM, also based on LibSVM software for SVM construction. Chiang et al. [78] also used a hierarchical SVM to classify the four categories of musical emotions. Its first SVM classifier was connected in cascade with the second third classifier, and the second third SVM classifier took parallel connections. The first SVM categorized the music samples into low arousal and high arousal groups, while the second and third SVMs again categorized the outputs of the previous SVM into positive and negative valence, thus achieving the purpose of recognizing the four categories of emotions. Hsu and Chen [102] proposed a structure consisting of a seven-layer SVM. Unlike general SVMs, the SVM in the article was based on a deep neural network (DNN), which was improved to a deep support vector machine (DSVM), and the output values of the previous layer of SVMs are part of the input vectors of the next layer of SVMs, and the structure could be repeated indefinitely. From the conclusion, it could be seen that as the number of SVM layers increased, the weighted average recall (WAR) also increased, and when the number of layers was increased

to seven, the WAR increased to 81.9%. We compared the multi-layer SVM structures described above Fig. 9

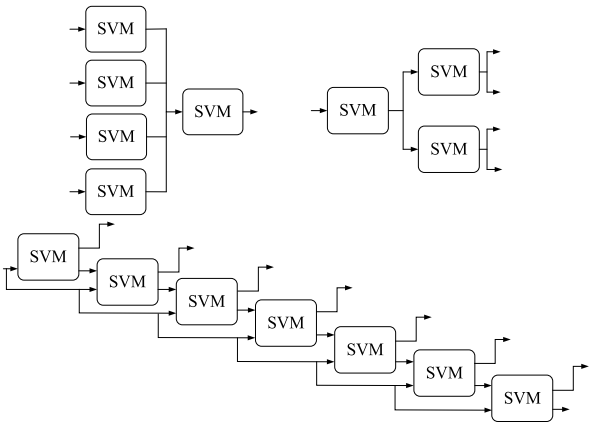


FIGURE 9. Schematic of the multilayer SVM structure described above, with the structure proposed by Chin et al. on the top left, Chiang et al. on the top right, and Hsu et al. on the bottom.

Random forest (RF) is a classifier that automatically selects a combination of multiple decision trees for classification, in addition to the feature selection mentioned above which can be done by evaluating the importance of each feature. Zhang et al. [60] used a random forest classifier based on the APM database for the experiment. They analyzed the accuracy of the obtained predictions with the actual results combined with the confusion matrix and finally concluded that the highest accuracy of 83.29% was obtained in the combination of the EEG features with the features extracted by OpenSMILE involved in the RF classifier. The multi-layer perceptron (MLP) is a feed-forward neural network model based on artificial neural networks and is a very simple classification model. It mainly consists of input, hidden, and output layers. And the MLP also uses a back-propagation supervised learning technique to train the network. Dutta et al. [38] MLP classifier for emotion classification of Assamese song database, the classifier with MFCC features involved in training has 93.75% accuracy in emotion recognition is much higher than the accuracy of 62.50% with the involvement of chroma features. Juthi et al. [77] trained four machine learning classifiers, artificial neural network (ANN), support vector machine (SVM), linear discriminative, and integrated learners for the task of music emotion recognition with separate inputs of the extracted features and compared their accuracy for emotion classification. The ANN achieved the highest level of accuracy, 75%. Shi et al. [103] used a system equipped with a two-layer Adaboost classifier to perform the music emotion classification task. The first layer input segmentation features (timbre features) and the second layer input super-segmentation features (rhythmic features), and the whole system consisted of three parts: feature extraction, hierarchical classifiers, and classification rules. After the combination of the segmentation features and the rhythmic features of the log-scale modulation frequency coefficients (LMFC) were introduced into the overall system, the average recognition accuracy of the four emotions could reach 92.8%. Bai et al. [107] compared and analyzed a number of classifiers such as support vector machine (SVM), k-nearest neighbor (KNN), neuro-fuzzy network classifier (NFNC), fuzzy KNN (FKNN), Bayesian classifiers, and linear discriminant analysis (LDA) based on classification problems. The final experimental results showed that SVM, FKNN, and LDA have the best performance in music emotion classification task, 82.7%, 83.0% and 80.4% respectively, and the accuracy of emotion recognition could reach more than 80%.

2) MODELS FOR SOLVING REGRESSION PROBLEMS

Support vector regression (SVR) is an application of support vector machine (SVM) to find the mapping function between inputs and outputs. Han et al. [104] proposed a music emotion recognition system based on SVR regressors. They first mapped the seven extracted music features onto a Thayer 2D emotion model, and then trained two SVR functions with the extracted features in polar coordinates, one calculating the distance from the origin to the emotion category in Thayer's polar coordinate system, and the other calculating the angle at which the emotion category is located. In addition, Han et al. separately trained SVM classifiers and gaussian mixture models (GMM) for experimental comparison. Finally, after transforming the Cartesian coordinate system to polar coordinate system, the accuracy of emotion recognition was greatly improved, with SVR and GMM having the highest accuracy of 94.55% and 92.73%, respectively. Yang et al. [105] used the extracted music features to train three regression algorithms, multiple linear regression (MLR), support vector regression (SVR), and AdaBoost.RT (BoostR). They compared MLR and BoostR as baseline methods and conducted an extensive performance study by selecting the data space, feature space, and regressions. The final results showed that SVR has the highest R^2 arousal and valence values in the principal component space and feature selection space (consisting of 18-dimensional top-selected features and 15-dimensional top-selected features), which are 58.3% and 28.1%, respectively. Deng et al. [106] introduced SVR to solve the music emotion regression problem by applying regression methods to predict the score value of each emotion class and calculating the accuracy of each of the eight emotions when they are used as the first dominant emotion and the second dominant emotion. They found that the proposed model is more sensitive to the recognition of two emotions, excitement and joy, which have accuracies of 0.845 and 0.839, respectively as the first dominant emotion, while it is less sensitive to the recognition of solemnity and dream, which have accuracies of only 0.476 and 0.512, respectively, as the first dominant emotion. Gaussian process (GP) is a Bayes non-parametric model, and GP regression is also an important method for dealing with the regression problem of music emotion recognition. Fukayama and Goto [108] used Emotion in Music Database as training data to construct GP regression and improved the accuracy of music emotion recognition by combining it

with adaptive aggregation, which has a higher performance improvement compare to GP regression with hybrid aggregation and GP regression alone Fig. 10. Chen et al. [109] proposed a deep Gaussian process based on Gaussian process regression, which is a learning method with a deep architecture that captures relationships between non-linear data. The results also show that the deep Gaussian process outperforms the traditional Gaussian process, demonstrating that the performance improvement potential of deep learning architectures is huge. Zhang et al. [30] created a novel dataset, PMEmo, and applied the widely used multiple linear regression (MLR) and support vector regression (SVR) as regressors to simulate the valence and arousal of emotions, and validated the feasibility of the created dataset between static and dynamic emotion recognition by evaluating their results.

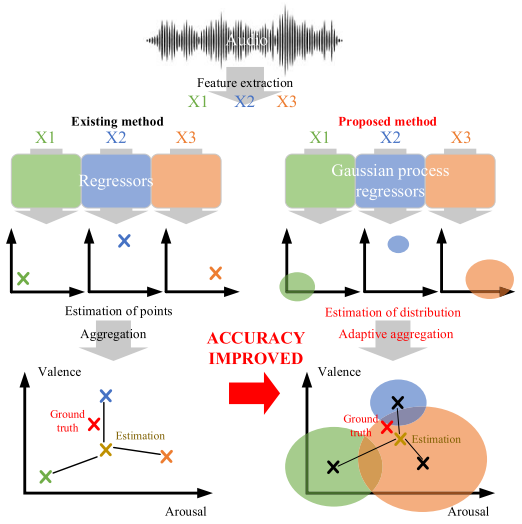


FIGURE 10. The fixed and adaptive aggregation of Gaussian process regressors for music emotion recognition.

B. DEEP LEARNING BASED MODELS

For music emotion recognition, to improve emotion recognition performance, either the original model is optimized and improved, or multi-modal features are added to synthesize music emotion recognition. Regarding multi-modal music emotion recognition, which has been applied in previous articles. For example, Zhang et al. [60] combined EEG and audio features to classify musical emotions, Ualibekova et al. [76] combined audio and textual features to classify musical emotions, and Rachman et al. [110] combined audio, text, and psycholinguistic features to classify musical emotions. However, there is less content related to multi-modal music emotion recognition based on traditional machine learning models. With the wide application of deep neural networks and the general popularity of cross-domain research in recent years, researchers have begun to conduct a lot of research on multi-modal music emotion recognition based on deep learning models. In these multi-modal music emotion recognition studies, audio features are combined with text features, visual features, and physiological features, among which multi-modal music emotion recognition based on audio and text features is very common.

1) UNIMODAL MUSIC EMOTION RECOGNITION

The general CNN architecture shown in Fig. 11 mainly consists of several functional layers [50]:

- *Convolutional layer composed of neurons:*
  They are connected from layer to layer through inputs or filters, and convolved through filters to obtain some features.
- *Batch normalization layer:*
  The main purpose is to normalize a small batch of input data to speed up network training and convergence.
- *Rectified linear unit layer:*
  Also known as the Relu layer, it is used to activate features and pass them to a layer.
- *Max pooling layer:*
  It performs non-linear down-sampling, which acts as a dimension reduction in the network structure of the CNN, and outputs the maximum value of it.
- *Full connection layer:*
  It takes the data output from the previous layers and predicts the emotion category for all songs.
- *Softmax layer:*
  As the classification layer of CNN, the music emotion classification result is output.

Chang et al. [21] proposed an end-to-end CNN architecture, which mainly consisted of the MS-SincResNet/MS-SincResNet model Fig. 12. A method called IIOF, namely intra- and inter-feature orthogonal fusion, was applied in this architecture for music emotion recognition research. They first used MS-SincNet/MS-SincNet to obtain 2D representations in the music clips, and obtained localized 2D representations through spatial average pooling (SAP) in the fourth layer (conv4) of ResNet-18, and global 2D feature representations through global average pooling (GAP) in the fifth layer (conv5). These feature representations were integrated by the IIOF module into the discriminative descriptors of the music emotion recognition task, and finally the music emotion annotations were estimated using a regressor. Unlike the network model proposed by Chang et al., Zhang et al. [14] proposed a music emotion recognition model based on an end-to-end CNN architecture called frequency embedded regularized network (FERN) Fig. 13. They added a frequency embedding module prior to the feature extraction task to improve the efficiency of the overall model feature extraction, and improved the ResNet-34 network that performed the feature extraction task by resizing its first two layers to retain more information. A modified criterion of (8) was used to adjust the ResNet-34 network radio frequency, the average pooling layer was replaced by a bipartite long short-term memory (Bi-LSTM) network layer to extract the sequential information, and finally output to the fully connected layer

for numerical prediction.

$$S_n = \begin{cases} 3 & if\ n \le \varepsilon \\ 1 & if\ n > \varepsilon \end{cases} \tag{8}$$

In (8), S denotes the core size and n denotes the number of remaining block layers, namely the criterion indicates that the number of layers is greater than ε and the core size is 1*1, and less than or equal to it is 3*3. Berardinis et al. [32] proposed a model called EMOMucs Fig. 14, which is based on the latest deep learning model for music source separation (MSS) Demucs. Demucs is based on U-Net convolutional neural network and adds Bi-LSTM between encoder and decoder, which solves the problem of irrecoverable information loss due to source mixing of conventional MSS. The overall model was modularized and divided into a total of four source-specific modules: vocals, drums, bass, and others, and finally MLP was applied to predict arousal and valence.

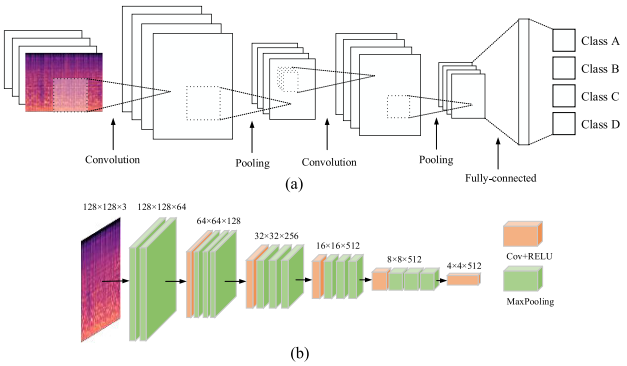


FIGURE 11. Schematic of the general convolutional neural network structure, where (a) is the CNN planar schematic and (b) is the CNN three-dimensional schematic.

Although the breakthroughs made by CNN in the field of music emotion recognition in recent years are evident to all, Tong [119] argued that CNN is suitable for processing two-dimensional data, and deep belief network (DBN), which are also deep learning models, are more suitable for processing one-dimensional music signals. Therefore, he proposed an improved DBN by adding a hidden layer node to each layer of the restricted Boltzmann machine (RBM) that constitutes the DBN, providing it with training data labels and fine-tuning all the weights, which can effectively improve the training accuracy of the model. Finally, the improved DBN combined with SVM classifier for music emotion classification model performs best compared to other deep learning classification models CNN+LSTM, LeNet, and ResNet. The average accuracy is 18.7% higher than the worst LeNet model, and even the CNN+LSTM model with the second highest average accuracy is 15.11% behind.

Long short-term memory (LSTM) network is a special architecture of recurrent neural network (RNN), which was formally proposed by Hochreiter and Schmidhuber [111] in 1997. It solves the problem that traditional RNNs are prone to gradient vanishing when dealing with long sequences. And

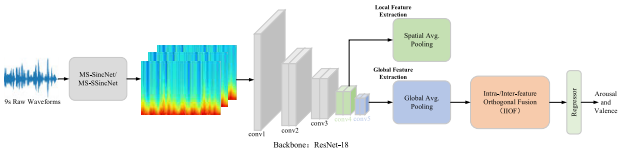


FIGURE 12. The CNN end-to-end model proposed by Chang et al.

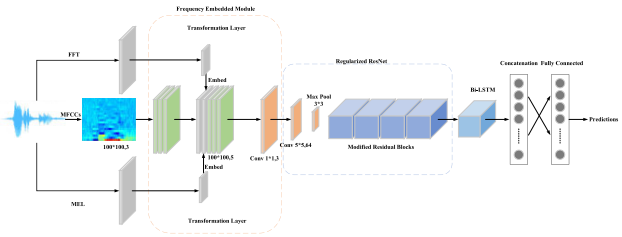


FIGURE 13. The CNN end-to-end model proposed by Zhang et al.

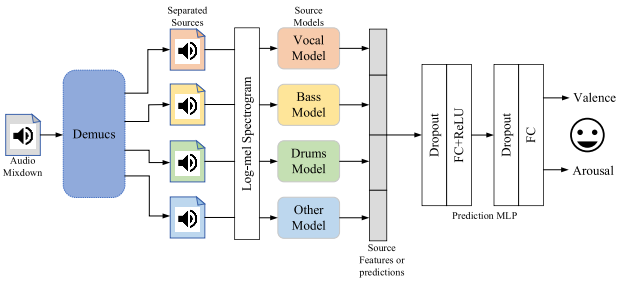


FIGURE 14. The EMOMucs model proposed by Berardinis et al.

through the cell state and gating mechanism, the LSTM network can better capture long term dependencies in sequence data, and it can also learn temporal features due to LSTM's sensitivity to time. Graves and Schmidhuber [112] proposed the bidirectional long short-term memory (Bi-LSTM) network in 2005, which is a network that improves on LSTM by adding a forward and a reverse LSTM network, respectively. This allows the model to take into account both future and past contextual information, and to process data better than LSTM in natural language processing tasks. In Grekow's [42] experiments, four variants of recurrent neural networks (RNNs) were used to construct regression equations to categorize musical emotions, each comparing features extracted by Marsyas with features extracted by Essentia. Among the comparisons, the linear regression and SMOreg algorithms were also added to the comparison term, and the final results showed that the RNN with two LSTM layers had better results in terms of both arousal and valence. In order to obtain better experimental results, Grekow improved the original experiments. He used a simple dense-layers neural network (NN) to build a pre-trained model to process the features extracted by Essentia into new features, and connected it to an RNN to achieve a relative improvement of R^2 6% in arousal regression and R^2 15% in valence regression for the best model. Jia [61] only recognized the emotion of the lyrics contained in the songs, and the CNN-LSTM network was

constructed by combining the CNN and LSTM networks. It also has the advantages of CNN for extracting local features and LSTM for ordering and connecting the extracted features. These features were extracted from Word2vec, while for TF-IDF the extracted bag-of-words model vectors are processed by a deep neural network (DNN) with a three-layer fully connected (FC) layer. In order to accurately extract fine-grained elementary target words and avoid low accuracy of emotion analysis, Jia also built a matching attention mechanism based on the original attention mechanism, and finally performed the emotion classification task by Softmax. Hizlisoy et al. [39] proposed a new model architecture CLDNN based on the original CNN and LSTM Fig. 15. They input the extracted MFCC and log MEL filter energy features into the CNN to obtain CNN-based features with the standard audio features extracted using the three publicly available feature extraction tools openSMILE, MIRtoolbox, and jAudio and input them into the model. LSTM+DNN was used as the classifier in the model to classify the music emotion, the LSTM layer mainly consisted of 200 hidden units and inputs the data to the two fully connected layers (DNN) which both consisted of 100 hidden units, and its input was activated by the Relu layer and outputs to the Softmax layer to get the final classification result. He and Ferguson [31] divided the music emotion recognition step into two general phases. In the first stage, a CNN-based autoencoder was used to compute the music feature representation, namely the unsupervised feature extraction part of the proposed model. And in the second stage, a bidirectional long short-term memory (Bi-LSTM) network was used to predict the emotions among the music fragments, namely the music emotion classifier part of the supervised learning. Du et al. [13] combined the more popular CNN and Bi-LSTM models for dynamic VA recognition of music emotion. They used two CNNs fed with Mel spectrogram and Cochleogram features, respectively, to train the weights from W1 to Wn, and their outputs were connected in a 32-dimensional fully-connected layer. After that, their outputs were input again to two 128-dimensional Bi-LSTMs respectively to learn the temporal information in the features, and the final regression layer was used to predict the final valence and arousal values Fig. 16. Due to the characteristics of CNN for processing two-dimensional data, such as spectrograms in music features, and RNN for processing time-dependent sequential data, Velankar et al. [113] combined their advantages to cascade and parallelize CNN and RNN networks, respectively, and proposed two network models, CRNN and parallel CNN-RNN. However, because their music dataset contained only a small dataset of 138 songs, the accuracy of these two types of models for emotion recognition was not very high, reaching only a little more than 50%. The model proposed by Gupta [45] also combined CNN and RNN. In his research, two attention based neural network models were proposed, an ACRNN (attention based convolutional recurrent neural network) with recurrent layers and an ACNN (attention based convolutional neural network with positional coding) without recurrent layers Fig. 17.

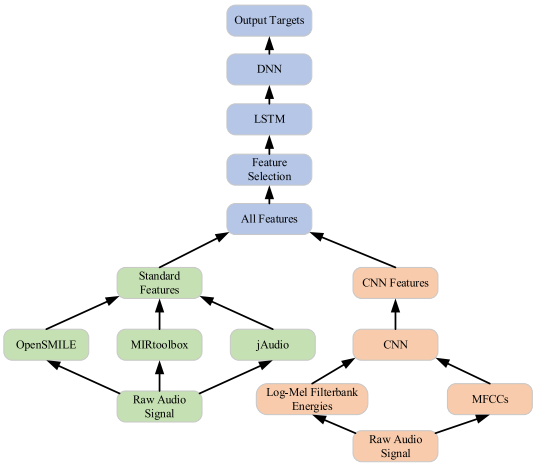


FIGURE 15. Schematic structure of the CLDNN model proposed by Hizlisoy et al.

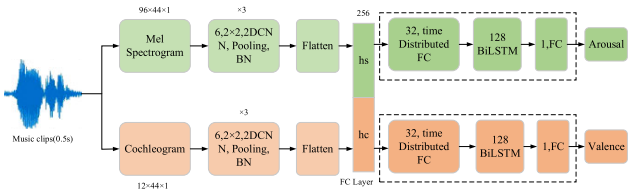


FIGURE 16. The model proposed by Du et al.

The ACNN and ACRNN were almost identical in terms of architecture, the only difference being that the ACNN learned the model weights and the attention weights separately by splitting the Time-Distributed-Flatten layer into two branches using positional coding. In addition, Gupta had included the Adam and AdamP optimizers among the proposed models and evaluated the $F1$-score, precision, recall, and accuracy of all the models. In the end, the ACNN model combined with the AdamP optimizer achieved the best evaluation results with an $F1$-score of 0.79 and an accuracy of 0.793, while the $F1$-score combined with Adam achieved 0.75. The music emotion recognition model proposed by Jia [89] improved the CNN structure by adding RNN to form the CRNN model Fig. 18. The overall model consisted of two classification models, CRNN and Bi-LSTM, where the input of CRNN was local and sequential feature extraction on the Mel spectrogram, and the input of Bi-LSTM was sequential feature extraction on the low-level audio features weighted by a combination of MFCC features and RP features. The outputs of both were connected to the full connectivity layer, where the emotions of the music were categorized using Softmax. In addition, Jia also improved the Softmax classifier by introducing a center loss function to control the center of the categories, which improved the differentiation between the categories and solved the shortcomings of the classifier's non-aggregation between the categories, and could better differentiate between similar emotions. The final proposed model can achieve a maximum recognition

accuracy of 92.06% with a loss value of only about 0.98. Dong et al. [22] proposed a bidirectional convolutional recursive sparse network (BCRSN) based on CNN and RNN. They used CNN instead of spectrogram input, and the first two hidden layers combined with LASSO regression feature selection method could adaptively learn salient emotional features (SII-ASF) containing sequence information from the spectrogram of music signals. In addition, they established bi-directional recursion for the neuronal connections of the first two hidden layers in the temporal order of each frame, and replaced the neurons in the bidirectional convolutional recursive feature mappings (BCRFMs) with LSTM modules to realize the successive emotion detection of audio files. Furthermore, to reduce the high computational complexity caused by numerical truth values, Dong et al. proposed a weighted hybrid binary representation (WHBR) approach. This method reduced the computational complexity by transforming the regression prediction process into a weighted combination of multiple binary classification problems and transforming the ground truth into a mixed binary vector. To ensure that the BCRFM was as sparse as possible and to avoid large parameters, they used L1 regularization as the objective function of the model.

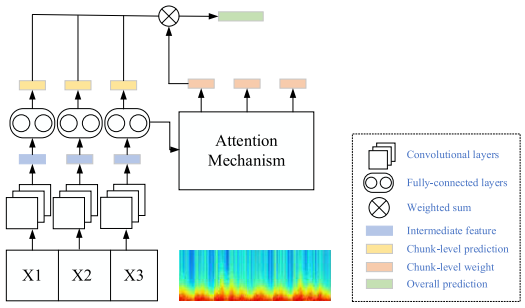


FIGURE 17. ACNN model architecture proposed by Gupta.

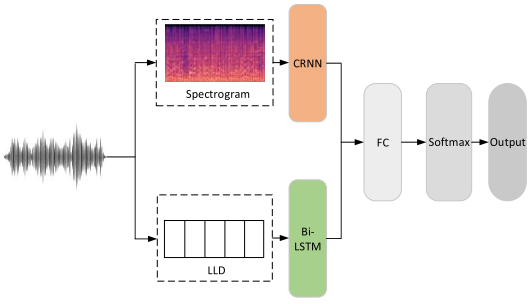


FIGURE 18. The model proposed by Jia et al.

We know that LSTM Fig. 19(a) is a special model among RNN models, and the gate control recursive unit GRU Fig. 19(b) is also a special model among RNN models. Similar to the function of LSTM, they are proposed to solve the problem of gradient vanishing in RNN models, but the difference between the structure of GRU and LSTM is that it is simpler, which makes the whole model easier to compute, and the computational efficiency is greatly improved. Han et al. [86] combined a one-dimensional convolutional neural network (CNN) with an optimized Inception module (Inception-GRU residual module with GRU gating unit) Fig. 20. Compared to ordinary CNN, it performs well in emotion classification tasks, with an accuracy of 84%. Ma and Zhou [40] conducted a study on emotion recognition of pure music for violin, and their proposed model was a model consisting of a convolutional neural network (CNN) with bidirectional gated recursive unit (Bi-GRU) and an attention mechanism (AM). They used the Bi-GRU and Attention characteristics to receive the music features output from the CNN and capture the temporal dependencies and important features in the music, and finally output them to the fully connected layer to perform the emotion classification task. Niu [95] proposed a bi-directional gated recursive unit (Bi-GRU) combined with a self-attention mechanism (SAM) model for music emotion classification. In contrast to the combination of LSTM and attention mechanism (AM), such an enhancement further improved the model fitting ability and reduced the computational time while increasing the computational efficiency. In addition, Niu proposed a LDA (latent-dirichlet al-location) topic scene classification model to find out the distribution of topics for each piece of music and the distribution of audio words in each topic. He significantly improved the accuracy of music recommendation by combining Bi-GRU+SAM with the LDA model to double categorize emotions in music.

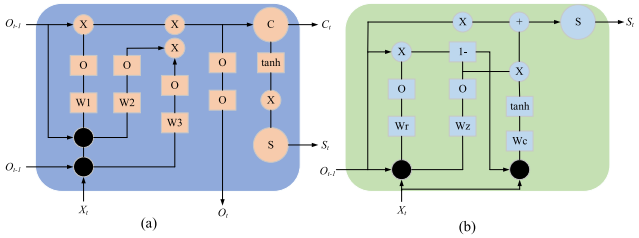


FIGURE 19. Comparison of the internal structure of LSTM and GRU, where (a) is the LSTM structure and (b) is the GRU structure.

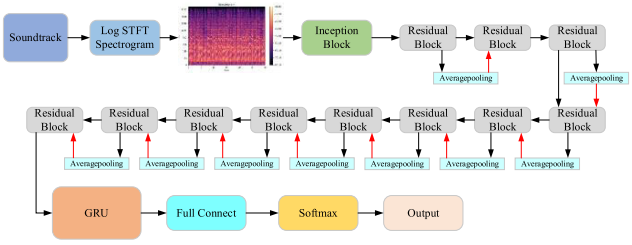


FIGURE 20. The model proposed by Han et al.

Huang et al. [114] proposed a generative adversarial network (GAN) in the feature extraction part among the music emotion recognition work. They solved the problem of GAN's inability to capture certain categories of music features by developing a GAN network model (DCGAN) based on a dual-channel attentional mechanism. The incorporation of an attention mechanism enabled the ability to

flexibly capture the relationship between local and global, and it could effectively extract local and global features of an image or sound. The DCGAN model included a feature attention model Fig. 21(a) and a channel attention model Fig. 21(b), which captured feature dependencies in feature space and channels, respectively. The dual-channel attention model Fig. 21(c) was the fusion of its outputs and the input feature maps to obtain a space E with global and local feature dependency information. The formula for E is given below:

$$E = \alpha P + \beta Q + X \tag{9}$$

In (9), P is the feature attention model output, Q is the channel attention model output, α and β are their hyperparameters, respectively, and X is the feature map. The dual-channel attention model proposed by Huang et al. had the highest music emotion recognition rate of 93.4% when comparing the feature attention model and the channel attention model alone. The overall network model DCGAN also has the highest recognition rate among the models CLSTM, RNN, GAN, and HTG for the music emotions sadness, happiness, quietness, loneliness, and missing. In particular, the recognition rate of two emotions, sadness and happiness, reached more than 90%. Yang [19] improved the BP neural network in the music emotion classification task by introducing the artificial bee colony (ABC) algorithm to adjust the weights and thresholds of the BP network and feedback the optimal solution to the BP network. This greatly improved the global search capability of the BP neural network while reducing probability of falling into a local optimal solution, resulting in faster convergence and a more stable BP neural network.

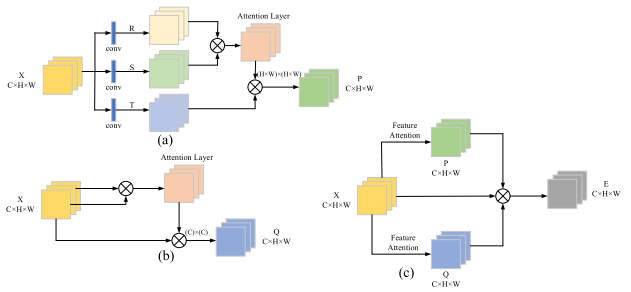
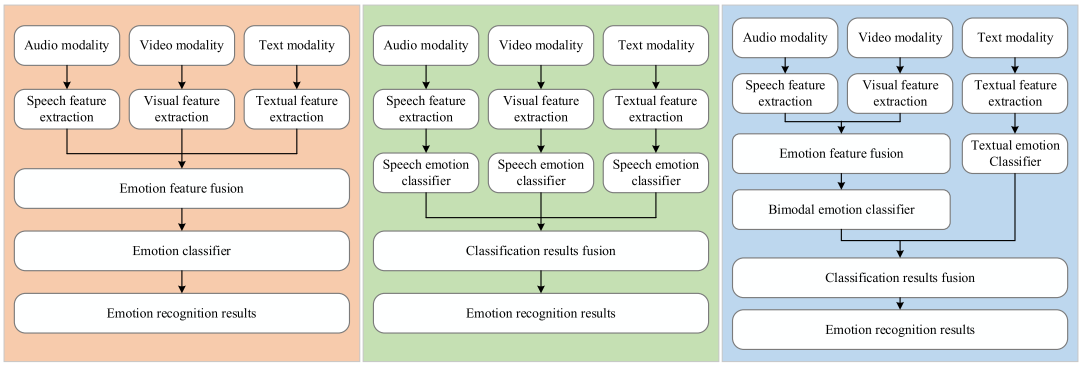


FIGURE 21. The model proposed by Huang et al. The figure shows (a) the feature attention model, (b) the channel attention model, and (c) the dual-channel attention model.

2) MULTI-MODAL MUSIC EMOTION RECOGNITION

Compared with unimodal music emotion recognition, multi-modal music emotion recognition can analyze the emotion shown by things from multiple angles, and the accuracy of emotion recognition is improved. However, multi-modal music emotion recognition means that multiple types of datasets must be processed and features extracted, which greatly increases the workload, and the constructed networks can become very complex. For multi-modal music emotion recognition, the current serious challenge is how to better integrate data from different sources. The fusion of multi-modal models can reduce the complexity of experiments while improving the accuracy of models for music emotion recognition tasks. General fusion methods for multi-modal music emotion recognition can be broadly categorized into three types, one is features layer fusion, another is decision layer fusion, and the last is hybrid layer fusion Fig. 22.

Chen [66] proposed a network structure with a two-layer LSTM with double channels Fig. 23, training the model with audio data containing music and video data containing dance movements and facial expressions, from which 43-dimensional features and 26-dimensional features were extracted, respectively. They used analytic hierarchy process (AHP) for model fusion at the decision layer of the network, comparing unimodal emotion recognition of audio and expression, the accuracy of multi-modal emotion recognition was improved by 7.9 percentage points and 13.2 percentage points, respectively. Among the multi-modal music emotion recognition experiments by Pandeya and Lee [115], they applied transfer learning to the well-known deep neural network to perform the music emotions classification task, which solved the problem of lack of training dataset. The networks used included a pre-trained 3D convolutional neural network (C3D) trained on the sport-1 M dataset, a pre-trained 3D Inception-V1 network (I3D) trained on the RGB ImageNet and Dynamics datasets, and an audio network of a pre-trained 2D convolutional neural network. In addition, Pandeya et al. also fine-tuned the audio network of the original one-dimensional convolutional neural network and performed a two-by-two video-audio network merging of the four networks mentioned above, with feature fusion at the decision layer and classification by Softmax, resulting in four multi-modal model architectures. The decision-level features of these four models are then combined into a single predictive model, resulting in a fifth model architecture, integrated multi-modal. In Liu et al.'s [116] network architecture for multi-modal music emotion recognition, an LSTM network is used as the main network model for audio. The overall audio network model consisted of an input layer that performed feature extraction, a two-layer LSTM hidden layer structure with 128 neurons in the first layer, 32 neurons in the second layer, and an output layer that contained 4 neurons. The Chinese pre-training model "BERTBase, Uncased" with a 12-layer transformer was used for the text network model. In addition, Liu et al. improved a subtask combined late fusion method (LFSM) based on a two-dimensional emotion model by proposing a new fusion approach. They took the next step based on the classification results of the audio and lyrics and introduced a neural network based on linearly weighted decision layer fusion. If the classification results belonged to the same category directly for decision layer fusion, did not belong to the lyrics through the emotion dictionary to analyze as a way to adjust the lyrics classification weights, and then with the adjusted audio content classification results for decision layer division, and finally obtained the music emotion classification results. In addition to the fusion performed at



(a) Feature-level fusion (b) Decision-level fusion (c) Hybrid-level fusion

FIGURE 22. The figure shows three fusion methods for multi-modal models, with the feature layer fusion method and the decision layer fusion method being the most common.

the decision layer, Delbouys et al. [25] also fused models at the feature layer. They fused the features extracted from the multi-modal dataset at the intermediate (second layer or later) and late (weighted output layer) stages, respectively. Comparison of end-to-end deep learning approaches revealed that deep learning-based intermediate fusion of multi-modal features achieved similar results to feature engineering and also outperformed two unimodal models (audio and text). Krols et al. [92] also verified the superiority of multi-modal emotion recognition by inputting features fused by separately feature layers into four different regression models and predicting valence and arousal on the Deezer emotion detection dataset.

Although feature fusion and decision fusion are widely used in multi-modal fusion methods, due to the large heterogeneous differences between different modal feature vectors, this leads to certain limitations of both feature fusion and decision fusion, which cannot transfer the information of modal features very effectively. Chen et al. [87] improved the CNN-LSTM classification model for the heterogeneity between 1D features (LLD in audio features and word frequency vectors in text features) and 2D features (spectrogram in audio features and word embedding in text features) among the unimodal, and proposed a multi-feature combinatorial network classifier consisting of 2D+CNN+LSTM and 1D+DNN. Aiming at the heterogeneity between different modalities, audio features and text features, Chen et al. also proposed a stacked integration learning method, which abandoned the traditional feature fusion and decision fusion. They trained the original sample features by audio multi-feature combination network classifier and lyrics multi-feature combination network classifier, combined the basic labeling results obtained from the training into a new dataset sample feature representation, and then input them to the sub-classifiers for learning and training, and finally output the integrated classification results. The overall network structure is shown in Fig. 24. Unlike the parallel approach of the multi-feature combinatorial network classifiers mentioned above, Yang et al. [117] used a CLDNN model architecture for the core algorithm under the music emotion recognition application. This was a network structure that connected CNN, LSTM, and DNN networks in a cascade fashion, where MFCC were feature extracted by CNN and output to LSTM+DNN network, which acted as classifiers to categorize music emotions.

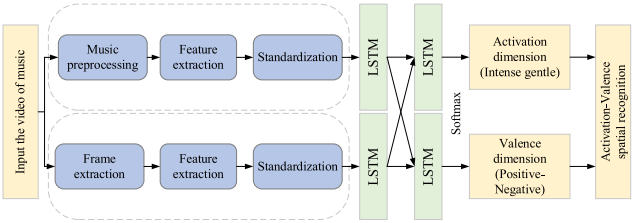


FIGURE 23. Multi-modal music emotion recognition proposed by Chen et al.

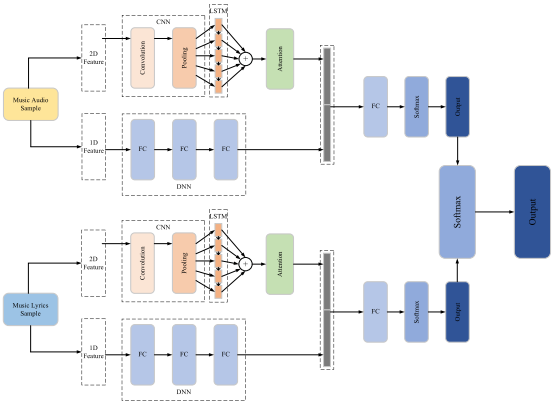


FIGURE 24. Structure of the overall multi-modal model network proposed by Chen et al.

Yang et al. [24] proposed a hierarchical model architecture called COSMIC Fig. 25, which was similar to the structure of a multi-modal music emotion classification model MMD-MII proposed by Wang et al. [20]. A cross-processing

module was mentioned in both their models Fig. 25, which consisted mainly of two Emotion-LSTM networks optimally enhanced on the basis of LSTM networks. The network combination received input from lyrics and audio pairs in addition to emotion vectors, which helped the model become adept at capturing and understanding the complex emotional nuances conveyed by music. This was a module specially designed for processing multi-modal information, and they implemented the interaction between audio and lyrics through a cross-processing module, which input the output emotion feature vectors into the next set of audio lyrics to ensure the consistency of the emotion representation.

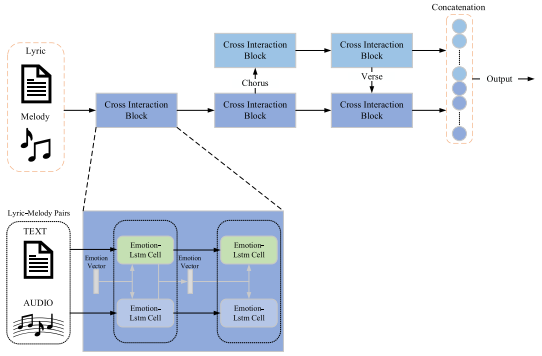


FIGURE 25. The structure of the COSMIC model proposed by Yang et al. and the internal structure of the cross-processing module in the structure.

Zhao et al. [26] classified music emotions based on audio, lyrics and contextual features among their multi-modal music emotion recognition experiments. They proposed a hierarchical network model structure Fig. 26, and introduced the cross-modal attention module (CMA) of the 8-head attention mechanism to hierarchically fuse the above three types of features in a certain order. By comparison with the model proposed by Delbouy and Pyrovolakis, state-of-the-art performance was demonstrated both in the R^2 score and in the $F1$-score. de Matos et al. [118] compared ResNet-18, AlexNet, VGG19, SqueezeNet, and DenseNet networks among multi-modal image and music emotion classifiers, and finally selected the ResNet-50 deep residual network with better classification results as the backbone of the proposed network structure to categorize emotions.

Sams and Zahra [99] proposed a model for multi-modal music emotion recognition for Indonesian songs, they used a CNN-LSTM model as classifier for audio. It combined the advantages of CNN, which was suitable for analyzing and processing image data, and LSTM, which could learn the temporal information in the features. It could handle images with time domain information such as songs and music videos very well. For text features, they used XLNet transformers as classifiers for text. This is an autoregressive language model, unlike the randomness of BERT, XLNet calculates all possible orders of each token to make predictions about words sequentially, so this approach allows more textual information to be obtained for the overall model, and is the

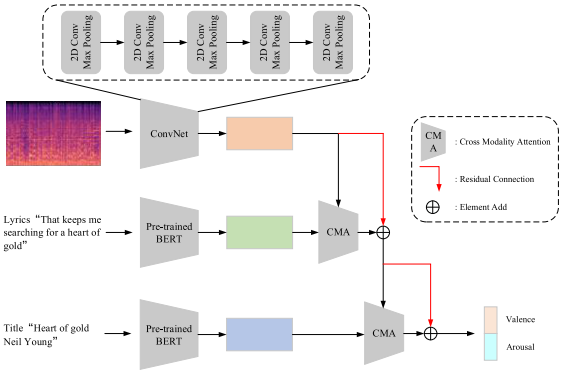


FIGURE 26. Diagram of the hierarchical network structure proposed by Zhao et al.

most advanced natural language processing (NLP) method at present. In addition, they applied the integration method of stacking in the music emotion classification task, which is mentioned in the study of Chen et al. above. Tong [90] combined knowledge distillation with transfer learning by using a teacher-student model Fig. 27 for knowledge distillation, which overcame the problem of less labeled data and unbalanced data. They also used an existing network architecture for music genre recognition and used its different stages of network parameters for the transfer learning process respectively. Regarding the music genre recognition network architecture was a multi-modal network architecture Fig. 28, which used a lightweight convolutional neural network for its audio model and discarded the classic networks such as ResNet, GoogLeNet, and VGGNet. The text in the textual model part contained information about the lyric of song, information about the title of the song, and the textual information about the title of album.

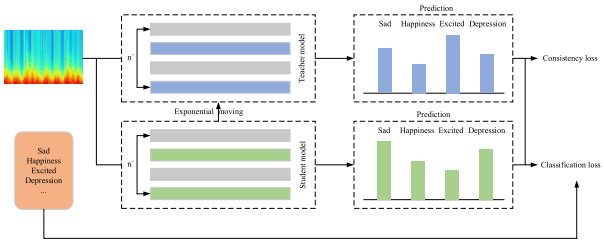


FIGURE 27. Teacher-student knowledge distillation model.

Thao et al. [51] proposed a deep neural network Fig. 29 (a) to classify whether music-video pairs match in terms of emotion, and the overall network structure consisted of three sub-networks. In the video sub-network, they pre-trained the SlowFast network with the Kinetics human action video dataset to extract spatial and temporal features, and added a fully-connected layer to dimension reduce them before inputting them to the projection head. In the music sub-network, they applied the pre-trained VGGish network on the AudioSet ontology to extract the features of each music clip of 0.98 s duration. And after dimension reduction of

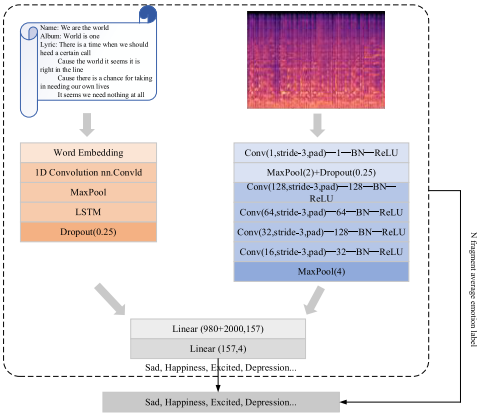


FIGURE 28. Architecture of a music genre recognition network for transfer learning by Tong et al.

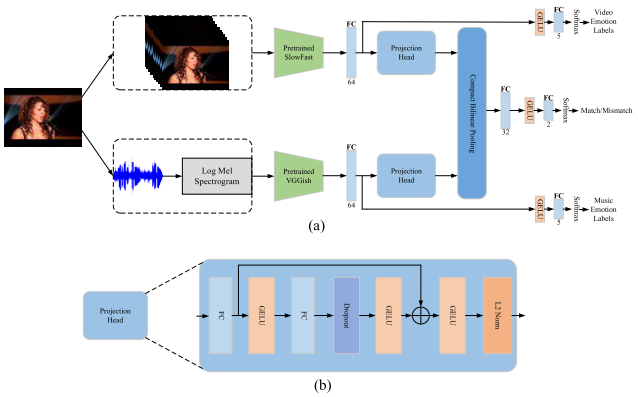


FIGURE 29. The figure shows the structure of the multi-modal music emotion recognition model proposed by Thao et al. and a zoom in on the internal structure of the projection head.

the features at the fully-connected layer, they used the same music projection head as the projection head structure of the video sub-network Fig. 29 (b) to embed the dimension reduced audio feature vectors into the common representation space. In the fusion sub-network, they obtained joint representations of video-audio through multi-modal compact bilinear pooling. They passed through a fully connected layer of 32 neurons, a GELU (gaussian error linear unit) activation function, and a fully connected layer of two neurons, and finally performed a binary classification in the Softmax layer to arrive at a match or mismatch in the emotion of the music-video pairs.

3) SUMMARY OF DEEP LEARNING BASED MODELS

From our research in music emotion recognition based on deep learning in recent years, we find that most deep learning models are based on convolutional neural networks or recurrent neural networks or a combination of the two. For example, as mentioned above, ResNet, U-Net, GAN, VGGNet, and Autoencoder are CNN-based neural networks; LSTM, Bi-LSTM, GRU, and Bi-GRU are RNN-based neural networks; and Demucs, CNN-LSTM, ACRNN, FERN, and CLDNN are CNN- and RNN-based neural networks. We also find that among the music emotion recognition models using deep neural networks as the backbone of the model, there are some differences in the researchers' choices of classifiers for music emotions. For example, in the above article, Chang et al. [21] used a regressor to predict VA values for music emotion recognition; Zhang et al. [14] predicted the VA value through the FC layer in the network; Berardinis et al. [32] predicted VA values by MLP; Jia [89] used Softmax to classify the emotions contained in music into four categories; Tong [119] classified discrete emotions in music by SVM. In order to understand these proposed deep learning models more intuitively, we present them in the form of a table in Table 3.

C. MODEL EVALUATION

R^2 is one of the more common model evaluation metrics in music emotion recognition, and is suitable for evaluating music emotion recognition tasks studied as regression problems, as it measures the fit of the regression model to the observed data as well as the accuracy of the prediction. Jandaghian et al. [91] calculated the mean square error and the average number of iterations for each emotion of the experimental results, and in addition, they used the R^2 statistic (coefficient of determination) (10) as another standard metric for model evaluation.

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{10}$$

In (10), RSS denotes the sum of squares of the residuals, TSS denotes the sum of squares, and R^2 coefficient of determination ranges from 0 to 1, with $R^2 = 1$ indicating that the model fits the data perfectly. The consistency correlation coefficient (CCC) (11), which combines the Pearson correlation coefficient (PCC), is also an important metric used to evaluate regression models, and it is widely used as a loss function (12). Zhang et al. [14] took two metrics, CCC and root-mean-square error (RMSE) to evaluate their proposed frequency-embedded regularized network (FERN) model. Comparing the models such as Bi-LSTM-RNN, Deep LSTM-RNN, and others, the FERN model showed excellent performance both in terms of valence and arousal. They obtained values of 0.328 and 0.134 in CCC and 0.418 and 0.121 in RMSE, respectively, which were the maximum and minimum values among the valence and arousal of the models for comparison.

$$\text{CCC}(X, Y) = \frac{\text{PCC}(X, Y)\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2 + (\mu_X, \mu_Y)^2} \tag{11}$$

$$\text{CCCLoss} = 1 - \text{CCC}(X, Y) \tag{12}$$

In (11) and (12), PCC is the PCC of the true value X and the prediction Y, σ_X and σ_Y denote the standard deviation of the two variables, μ_X and μ_Y denote the mean of the two variables, respectively. The value of CCC ranges from -1 to 1, then the value of CCCLoss ranges from 0 to 2.

TABLE 3. Deep learning based emotion recognition model for music.

Article	Datasets	Emotion Models	Preprocessing	Extracted Features	Network Models	Experimental Results (Partial)
Chang *et al.* [21]	DEAM and PMEmo	VA model	Data augmentation and music clip layer normalization	Spectrogram features	MS-SincResNet/ MS-SSincResNet	MS-SincResNet + IIOF (R=3) had an arousal accuracy of 86.7 and an F1-score of 91.17
Berardinis *et al.* [32]	PMEmo	VA model	Padding operations for tracks shorter than 20 s	Log-Mel spectrogram features	U-Net, Bi-LSTM, and MLP	The RMSE of the VA for the C1D model trained under mix-down was 0.2600 and 0.2444, and the R2 of the VA was 0.3489 and 0.5573
Tong [119]	FMA-Small	Sad, Happy, Quiet, Passionate, Romantic, and Thrilling	Vocal separation and fine-grained segmentation	Pitch frequency and band energy distribution features	Improved DBN and SVM	The accuracy of Sad, Happy, Quiet, Passionate, Romantic, and Thrilling, was 79.58%, 86.26%, 78.26%, 88.31%, 75.65%, and 82.64%, respectively
Grekow [42]	GTZAN	VA model	Signal framing	Features extracted by Marsyas and Essentia	LSTM	The RNN4 model trained on the features extracted by Marsyas had R2 and MAE of 0.67 and 0.12 for its arousal and 0.17 and 0.15 for its valence.
Jia et al. [61]	Own	Happy, Sad, Calm, and Healing	Removing redundant information, converts text to a digital vector, and sets the text vector dimension to 100 dimensions	Features extracted by TF-IDF and Word 2 vec	CNN and LSTM	The accuracy of Happy, Sad, Calm, and Healing was 0.903, 0.864, 0.809, and 0.816 respectively and their average accuracy was 0.848.
Hizlisoy et al. [39]	Turkish emotional music (TEM)	——	Feature selection, signal framing, and windowing	Timbre, energy, log-Mel filter-bank energies, and MFCC features	CNN, LSTM, and DNN	Compared to k-NN, SVM, and RF classifiers, LSTM + DNN classifier improved music emotion recognition accuracy by 1.61, 1.61, and 3.23 percentage points, respectively.
He et al. [31]	PMEmo and AllMusic	VA model	Signal framing and data normalization	Log-mel spectrogram features	Autoencoder and Bi-LSTM	Proposed model had a valence accuracy of 79.01 and an F1-score of 83.2; and arousal accuracy of 83.62 and an F1-score of 86.52.
Du et al. [13]	1000 Songs	VA model	Signal framing	Mel spectrogram and Cochleogram features	CNN and Bi-LSTM	The RMSE values of the proposed model were 0.06 ± 0.04 and 0.07 ± 0.05 for valence and arousal, respectively.
Velankar et al. [113]	Own	Devotional, Sad, and Romantic	Windowing	Mel spectrogram features	CNN and RNN	The CRNN model performed best for sad emotion recognition with values of 0.65, 0.69, and 0.67 for precision, recall, and F1-score respectively.
Gupta [45]	4Q audio emotion and Bi-modal emotion	——	Signal framing and deletion of segments of less than 8 s	Spectrogram features	CNN and RNN	The ACNN model outperformed the other models with an F1-score of 0.79 using the AdamP optimizer.
Jia [89]	Own	Anger, Happy, Relaxation, and Sad	Detecting mute frames in music signals with a voice activity detection	LLD and Mel spectrogram features	CNN, RNN, and Bi-LSTM	The proposed model had a recognition accuracy of 92.06% and a loss function of about 0.98.

TABLE 3. *(Continued.)* **Deep learning based emotion recognition model for music.**

Dong et al. [22]	DEAM and Mood Swings Turk (MTurk)	VA model	Windowing, features selection, dimension reduction, and signal framing	Spectrogram features	CNN and RNN	The RMSE, PCC, and CCC of proposed BCRSN model on the DEAM dataset for valence and arousal were 0.195 ± 0.114, 0.455 ± 0.230, 0.261 ± 0.257, and 0.124 ± 0.101, 0.505 ± 0.215, and 0.413 ± 0.100, respectively.
Han et al. [86]	Soundtra-ck	Happy, Angry, Sad, and Neutral	Signal framing, windowing, signal completion, data normalization	Spectrogram features	CNN and GRU	Deep learning model music emotion recognition accuracy of 84.23% for 1DCNN with Inception-GRU.
Ma et al. [40]	VioMusic	VA model	Data normalization	Mel spectrogram and LLD features	CNN and Bi-GRU	The MAE of the proposed CBA model was 0.124 and 0.129, respectively.
Niu [95]	GTZAN and ISMIR2004	Happy, Refreshing, Relaxing, and Sad	Feature dimension reduction and data normalization	Rhythmic and spectrum features	Bi-GRU	Bi-GRU was able to correctly recognize happy and sad emotional music with an accuracy of 79% and 81.01%, respectively.
Huang et al. [114]	Own	Sad, Happy, Quiet, Lonely, and Miss	——	MFCC features	GAN	The accuracy of the proposed GAN network model exceeded 87 %.
Yang [19]	MEM	VA model	——	Features such as short-term energy, short-term autocorrelation function, and spectrum	Improved BP	The difference between the MAE and RMSE values of the proposed model on the dimensions of valence and arousal was 0.284 and 0.0256, respectively.
Chen [66]	Bi-modal emotion	VA model and 6 classes of discrete emotions	Windowing and data normalization	Facial expression, MFCC, and Fbank features	LSTM	The proposed LSTM model achieved an accuracy of 77.9 %.
Pandeya *et al.* [115]	Own	Excitation, Fear, Neutral, Relaxation, Sad, and Tension	Data zero-padding, data normalization	Mel spectrogram and visual features	CNN	The prediction model integrating all multi-modal structures achieved an accuracy of 88.56% with an F1-score of 0.88, and an area under the curve (AUC) score of 0.987.
Liu *et al.* [116]	Music Mood Classification	Anxious, Exuberance, Depression, and Contentment	Capture the middle 30 seconds of the music clip and lyrics into the appropriate format.	MFCC, spectrum centroid, and frequency-band energy distribution features	LSTM and BERT	The optimized fusion algorithm improved over the linear weighted multi-modal fusion algorithm and linear least squares fusion algorithm by 5.77% and 4.03%, respectively.
Delbouys *et al.* [25]	MSD	VA model	Batch normalization, windowing, data augmentation, extraction of 30 s long audio clips	Mel spectrogram and word embedding features	CNN and LSTM	The R2 score for the best weighted average of classical and deep learning methods for multi-modal valence prediction was 0.243.
Chen *et al.* [87]	MSD	Angry, Happy, Relaxed, Sad, and Average	Fine-grained segmentation and vocal separation	Spectrogram, LLD, HSF, Word2vec, and Chi-squared test features	CNN, LSTM, Bi-LSTM, and DNN	The proposed multi-feature combination network classifier achieved 68% audio classification accuracy, 74% lyrics classification accuracy, and 78% multi-modal average classification accuracy.
Yang *et al.* [117]	PMEmo	VA model	——	MFCC features	CNN, LSTM and DNN	The accuracy of the proposed model was 90%.

TABLE 3. *(Continued.)* **Deep learning based emotion recognition model for music.**

Yang *et al.* [24]	Own	Happy, Sad, Quiet, and Healing, VA model	Source separation, lyric-melody alignment, lyric polarity intensity annotation, verse-chorus annotation	emotion and other features of the melody , text features	Improved LSTM and MLP	The values of the proposed COSMIC model for precision, recall, accuracy and F1-score were 0.4870, 0.4870, 0.4839, and 0.4841 respectively.
Zhao *et al.* [26]	MSD and MoodyLyrics	VA model	Windowing and batch normalization	Mel spectrogram, lyrics, and context features	CNN and BERT	The mean value of R2 of VA for the proposed model was 0.309 and the mean value of F1-score was 96.14%.
Sams *et al.* [99]	Own	——	Clean and trim the blank sound, clean and lowercase the lyrics	Mel spectrogram and lyrics features	XLNet, ANN, CNN, and LSTM	The proposed multi-modal model could achieve up to 85.11% recognition accuracy.
Tong [90]	Own	Sad, Happiness, Excited, and Depression	Detects silent frames in music signals, slices character words by special characters, data augmentation, and signal framing	Mel spectrogram and text features	CNN	The accuracy of the proposed multi-modal knowledge refinement model incorporating a genre porting algorithm was 81.3%.
Zhang *et al.* [14]	1000 Songs	VA model	Windowing and framing of signals	Mel spectrogram features	Improved ResNet and Bi-LSTM	The CCC and RMSE of the proposed FERN model were 0.328 and 0.134 for valence and 0.418 and 0.121 for arousal
Thao *et al.* [51]	EmoMV	——	Feature dimensionality reduction	Log Mel spectrogram and visual features	VGGish and SlowFast	The proposed model had the highest values of classification accuracy, F1-score, and AUC in EmoMV-A dataset, which were 79.03%, 0.80, and 0.87, respectively.

The closer the value of CCC is to 1, the stronger the consistency between the two variables is indicated. In the article by Koh and Dubnov [46], in addition to calculating the R^2 value of the proposed OpenL3 embedding on arousal and valence annotations on model evaluation based on the Emotion in Music dataset to compare to Choi et al.'s value of R^2. They also compared the F1-score of the classification results in Malheiro et al.'s study on another dataset, and the OpenL3 embedding of the SVM classifier outperformed Malheiro et al.'s F1-score by about 16%. The F1-score here is also a widely used model evaluation metric, and unlike R^2, the F1-score applies to classification models, which is the harmonic mean of precision (14) and recall (15), that is:

$$\mathrm{F1} = 2 \times \frac{\mathrm{TP}}{2 \times \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{13}$$

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{14}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{15}$$

In (13), (14), and (15), TP is true positive, FP is false positive, FN is false negative, TN is true negative, and the F1-score takes a value ranging from 0 to 1, with larger values indicating better model performance. The confusion matrix Fig. 30 can be thought of as a visual representation of the F1-score, and by looking at the confusion matrix and calculating it, we can accurately derive the F1-score.

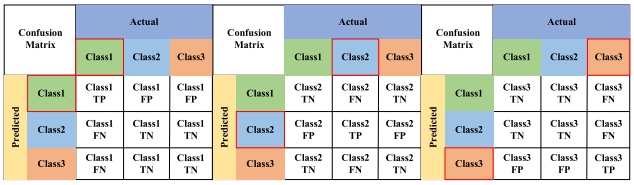


FIGURE 30. Visual representation of F1-score - confusion matrix.

It is worth noting that in Yang's [19] experiments, he used BP algorithm to compare the results of six combinations of features and analyzed which combination makes the best effect of emotion recognition by using to the accuracy of the model evaluation. Finally, he concluded that the feature combination consisting of short-time energy, short-time average amplitude, short-time autocorrelation function, short-time over-zero rate, frequency spectrum, and amplitude spectrum has a better recognition effect, and the accuracy rate can

reach 83.83%. The formula for calculating the accuracy is as follows:

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{16}$$

The symbols in the equation are expressed as above for the F1-score. However, when the positive and negative samples are unbalanced, the evaluation metric of accuracy Acc is highly flawed and does not reflect the model's predictive ability for a few categories of samples. Among the regression models proposed by Xia et al. [88], their study used three different regression methods for the music emotion recognition task. Considering that the mean error causes positive and negative errors to offset each other, resulting in a small error, they chose the mean absolute error (MAE) (17) to evaluate their proposed model.

$$\text{MAE} = \frac{1}{m}\sum_{i-1}^{m}|x_i - y_i| \tag{17}$$

In (17), m is the number of samples, x_i is the actual value of the samples, y_i is the predicted value of the samples, and the smaller the MAE value, the higher the predictive accuracy of the model is indicated. Ma et al. [40] used the Pearson correlation coefficient r (18), in addition to calculating the MAE for model evaluation, in a violin music emotion classification task. Among the results of the comparison of the models, their proposed model has the highest Pearson correlation coefficients r of 0.524 and 0.576 for model valence and arousal, and the lowest mean absolute error (MAE) of 0.124 and 0.129, with the optimal model.

$$r = \rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \tag{18}$$

In (18), $\text{cov}(X, Y)$ denotes the covariance of X, Y and σ_X, σ_Y denote their respective standard deviations, and the Peirson correlation coefficient takes values between -1 and 1: $r = 1$ for a perfect positive correlation, $r = -1$ for a perfect negative correlation, and $r = 0$ for no linear relationship. As mentioned earlier, in classification tasks, especially when the dataset categories are unbalanced, relying solely on accuracy may be misleading, so researchers often take multiple model evaluations to comprehensively evaluate the proposed model, and ROC-AUC is one of them, as in [115], which gave AUC scores to the models under the ROC curve Fig. 31. In a comparison experiment between the four multi-modal model architectures generated by Pandeya et al. [115] through combination and the integrated multi-modal model with parallel decision making, they used accuracy, $F1$-score, and ROC-AUC to evaluate the C3D+1D Music CNN, I3D+1D Music CNN, C3D+2D Music CNN, I3D+2D Music CNN, and five integrated multi-modal models. The final proposed integrated multi-modal prediction model achieved the best results with 88.56% accuracy, $F1$-score of 0.88, and area under the ROC curve (AUC) score of 0.987.

Thao et al. [51] compared the model performance evaluation of the proposed model with the baseline model on the EmoMV-A, EmoMV-B and EmoMV-C datasets they created, using three evaluation metrics: accuracy, $F1$-score, and AUC-score, all of which achieved better results. Among them, their proposed model performed best on the EmoMV-A dataset with an accuracy of 79.03, an $F1$-score of 0.80, and an AUC-score of 0.87. What we speculate that the EmoMV-A dataset contained the largest number of music video clips, and thus the model was trained the best. Xu et al. [48] evaluated the model performance of the proposed multi-modal music emotion recognition modeling approach for audio and lyrics through a 10-fold cross-validation method in a regression task study of the effect of lyrics features on music arousal perception and valence perception based on LIWC. Each regressor prediction accuracy was measured by the above mentioned R^2 in addition to the root mean square error RMSE (19) to measure the deviation between the experimental predicted and true values.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \tag{19}$$

The symbols in equation (19) are the same as in the MAE, the smaller the RMSE value, the higher the predictive accuracy of the model.

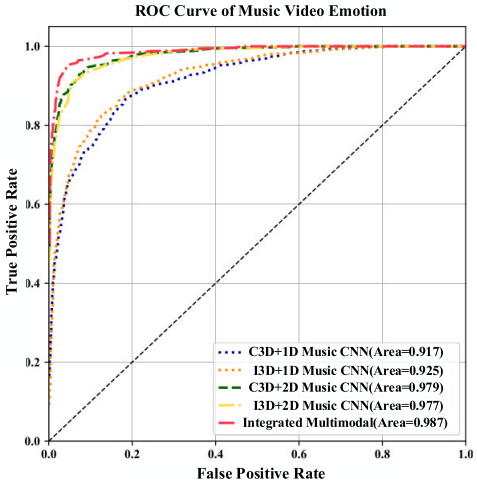


FIGURE 31. ROC-AUC score plot from Pandeya et al.

D. APPLICATIONS OF MUSIC EMOTION RECOGNITION

Lucia-Mulas et al. [85] combined the constant-Q-transform spectrogram, which best represents the relationship between human-perceived musical pitches, with a convolutional neural network. Their proposed model was applied to automatically identify the emotional intent of different segments in a movie soundtrack, which helps optimize automatic subtitling and stylistic adaptation. de Matos et al. [118] proposed a multi-modal hyperlapse method based on video and music emotion alignment, which recognized the emotion of video and music respectively, and created the emotion curves of both to align the emotions in video and music. They found

the best matching path between video and music by selecting optimal paths, which is applied in the current popular short video editing, so that the emotions in the edited video and background music can be better combined to enhance the viewers' emotional experience. Niu [95] analyzed the current state of the art in music classification and recommendation, and found that conventional music recommendation systems usually classify music according to language, music style, thematic scene, and chronology. However, there were difficulties in recognizing music emotion categories, and the accuracy rate of emotion recognition was too low. Therefore, he proposed a music emotion recognition model using gated recursive unit networks and multi-feature extraction, combined with a music recommendation model framework for topic classification, which improved the performance of music recommendation systems and could more accurately satisfy user needs. Yang et al. [117] used the music emotion recognition model CLDNN as the core of the design system, and designed a system that can search for playing music according to a person's password, and recognize the emotions among the music played. Then according to the recognized 12 categories of emotions: excitement, happiness, pleasure, relaxation, quietness, calmness, fatigue, boredom, sadness, anxiety, anger, and pain, it was passed to the light module and made 12 kinds of corresponding light mode feedback system. The system included a voice interaction module to receive voice commands, a WS2812 light module to provide light feedback on the emotions received, and a CLDNN algorithm module that was the core module of the system.

VIII. SUMMARY OF CURRENT CHALLENGES AND FUTURE RESEARCH

Music emotion recognition is inter-disciplinary research with a wide range of applications in various fields. Although the performance of music emotion recognition has continued to improve with the development of deep learning, it is generally in the rising stage and still faces many challenges at present:

1) Unlike speech emotion recognition datasets, music emotion recognition datasets, despite the increase in recent years, are still deficient due to music copyrights, the difficulty of creating homemade datasets, and the diversity of experiments. In multi-modal music emotion recognition research, it is also due to the lack of datasets based on visual and physiological features that fewer studies have been done on music emotion recognition using these multi-modal features. How to access to high quality music datasets and how to fill the gaps in music datasets with multi-modal features are still among the problems to be solved.
2) The current common approach to music feature extraction is still to extract the underlying physical features of the music and to analyze and process these features, but the connection between the underlying features and the higher-level emotions is still limited. Existing music features are unable to effectively express emotion-related information, and the establishment of a reasonable music feature analysis model is necessary for the current field research.
3) Due to cultural and linguistic differences in music, the models, datasets, and emotion categorization criteria used by researchers for music emotion recognition vary among different geographic regions. This has also led to inconsistent music emotion recognition models and model evaluation metrics that do not accurately measure model performance. Therefore, the generalizability of the model in multi-cultural and multilingual environments and the availability of an accepted and standardized model evaluation metric become a challenge.

Through the review of related research in the field of music emotion recognition, for future research, we believe that we should focus on the establishment of music datasets and the optimization of music emotion recognition models. For the creation of music datasets, large-scale cross-cultural datasets should be created in the future and complemented with datasets that include textual, visual and physiological features. By combining the category emotion model and the dimension emotion model in the emotion annotation of the dataset makes the emotion categories more centralized and the differentiation between different emotion categories greater, which facilitates the unification of the emotion categories involved in the performance evaluation of the model, thus promoting the high-quality construction of the music emotion dataset. Deep learning has better feature learning capabilities and a more essential description of the data compared to machine learning. Therefore, the future music emotion recognition model should mainly apply deep learning methods to effectively improve the recognition performance, and optimize and improve the original deep neural network as well as combine other advanced methods, which is expected to further improve the recognition performance. In terms of modality, it is also possible for future research to try to combine a variety of different modal features or different levels of feature fusion to achieve a more ideal music emotion recognition effect.

REFERENCES

[1] L. Perlovsky, "Musical emotions: Functions, origins, evolution," *Phys. Life Rev.*, vol. 7, no. 1, pp. 2–27, Mar. 2010.
[2] J. J. Fang, "Prehistoric music archaeology of the Yellow River basin and the origin of Chinese musical civilization," *J. Musical Res.*, vol. 5, no. 2, pp. 5–15, 2024.
[3] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," in *Proc. ISMIR*, vol. 86, 2010, pp. 937–952.
[4] T. Eerola and J. K. Vuoskoski, "A review of music and emotion studies: Approaches, emotion models, and stimuli," *Music Perception*, vol. 30, no. 3, pp. 307–340, Feb. 2013.
[5] Y. H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 3, pp. 1–30, 2012.
[6] X. Yang, Y. Dong, and J. Li, "Review of data features-based music emotion recognition methods," *Multimedia Syst.*, vol. 24, no. 4, pp. 365–389, Jul. 2018.
[7] R. Panda, R. Malheiro, and R. P. Paiva, "Audio features for music emotion recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 68–88, Jan. 2023.

[8] D. Han, Y. Kong, J. Han, and G. Wang, "A survey of music emotion recognition," *Frontiers Comput. Sci.*, vol. 16, no. 6, Dec. 2022, Art. no. 166335.

[9] X. Cui, Y. Wu, J. Wu, Z. You, J. Xiahou, and M. Ouyang, "A review: Music-emotion recognition and analysis based on EEG signals," *Frontiers Neuroinform.*, vol. 16, Oct. 2022, Art. no. 997282.

[10] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 songs for emotional analysis of music," in *Proc. 2nd ACM Int. Workshop Crowdsourcing Multimedia*, Oct. 2013, pp. 1–6.

[11] T. Liu, L. Han, L. Ma, and D. Guo, "Audio-based deep music emotion recognition," in *AIP Conf. Proc.*, vol. 1967, no. 1, 2018.

[12] B. Xie, J. C. Kim, and C. H. Park, "Musical emotion recognition with spectral feature extraction based on a sinusoidal model with model-based and deep-learning approaches," *Appl. Sci.*, vol. 10, no. 3, p. 902, Jan. 2020.

[13] P. Du, X. Li, and Y. Gao, "Dynamic music emotion recognition based on CNN-BiLSTM," in *Proc. IEEE 5th Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, Jun. 2020, pp. 1372–1376.

[14] M. Zhang, Y. Zhu, N. Ge, Y. Zhu, T. Feng, and W. Zhang, "Frequency embedded regularization network for continuous music emotion recognition," in *Proc. IEEE Int. Conf. Prog. Informat. Comput. (PIC)*, Dec. 2021, pp. 426–431.

[15] Z. Zhong, H. Wang, B. Su, M. Liu, and D. Pei, "Music emotion recognition fusion on CNN-BiLSTM and self-attention model," *Comput. Eng. Appl.*, vol. 59, no. 10, pp. 94–103, 2023.

[16] Y.-A. Chen, Y.-H. Yang, J.-C. Wang, and H. Chen, "The AMG1608 dataset for music emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 693–697.

[17] Y. Shu and G. Xu, "Emotion recognition from music enhanced by domain knowledge," in *Proc. 16th Pacific Rim Int. Conf. Artif. Intell.*, Yanuca Island, Fiji. Cham, Switzerland: Springer, 2019, pp. 121–134.

[18] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Developing a benchmark for emotional analysis of music," *PLoS ONE*, vol. 12, no. 3, Mar. 2017, Art. no. e0173392.

[19] J. Yang, "A novel music emotion recognition model using neural network technology," *Frontiers Psychol.*, vol. 12, Sep. 2021, Art. no. 760060.

[20] J. Wang, A. Sharifi, T. R. Gadekallu, and A. Shankar, "MMD-MII model: A multilayered analysis and multimodal integration interaction approach revolutionizing music emotion classification," *Int. J. Comput. Intell. Syst.*, vol. 17, no. 1, p. 99, Apr. 2024.

[21] P.-C. Chang, Y.-S. Chen, and C.-H. Lee, "IIOF: Intra- and inter-feature orthogonal fusion of local and global features for music emotion recognition," *Pattern Recognit.*, vol. 148, Apr. 2024, Art. no. 110200.

[22] Y. Dong, X. Yang, X. Zhao, and J. Li, "Bidirectional convolutional recurrent sparse network (BCRSN): An efficient model for music emotion recognition," *IEEE Trans. Multimedia*, vol. 21, no. 12, pp. 3150–3163, Dec. 2019.

[23] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proc. 12th Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, San Diego, CA, USA, 2011, pp. 591–596.

[24] L. Yang, Z. Shen, J. Zeng, X. Luo, and H. Lin, "COSMIC: Music emotion recognition combining structure analysis and modal interaction," *Multimedia Tools Appl.*, vol. 83, no. 5, pp. 12519–12534, Jul. 2023.

[25] R. Delbouys, R. Hennequin, F. Piccoli, J. Royo-Letelier, and M. Moussallam, "Music mood detection based on audio and lyrics with deep neural net," 2018, *arXiv:1809.07276*.

[26] J. Zhao, G. Ru, Y. Yu, Y. Wu, D. Li, and W. Li, "Multimodal music emotion recognition with hierarchical cross-modal attention network," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2022, pp. 1–6.

[27] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 2, pp. 467–476, Feb. 2008.

[28] S.-Y. Wang, J.-C. Wang, Y.-H. Yang, and H.-M. Wang, "Towards time-varying music auto-tagging based on CAL500 expansion," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2014, pp. 1–6.

[29] X. Liu, Q. Chen, X. Wu, Y. Liu, and Y. Liu, "CNN based music emotion classification," 2017, *arXiv:1704.05665*.

[30] K. Zhang, H. Zhang, S. Li, C. Yang, and L. Sun, "The PMEmo dataset for music emotion recognition," in *Proc. ACM Int. Conf. Multimedia Retr.*, Jun. 2018, pp. 135–142.

[31] N. He and S. Ferguson, "Music emotion recognition based on segment-level two-stage learning," *Int. J. Multimedia Inf. Retr.*, vol. 11, no. 3, pp. 383–394, Sep. 2022.

[32] J. De Berardinis, A. Cangelosi, and E. Coutinho, "The multiple voices of musical emotions: Source separation for improving music emotion recognition models and their interpretability," in *Proc. ISMIR*, 2020, pp. 310–317.

[33] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012.

[34] R. E. S. Panda, R. Malheiro, B. Rocha, A. P. Oliveira, and R. P. Paiva, "Multi-modal music emotion recognition: A new dataset, methodology and comparative analysis," in *Proc. 10th Int. Symp. Comput. Music Multidisciplinary Res.*, 2013, pp. 570–582.

[35] I. Dufour and G. Tzanetakis, "Using circular models to improve music emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 12, no. 3, pp. 666–681, Jul. 2021.

[36] B. K. Baniya and J. Lee, "Rough set-based approach for automatic emotion classification of music," *J. Inf. Process. Syst.*, vol. 13, no. 2, pp. 400–416, 2017.

[37] B. Jeon, C. Kim, A. Kim, D. Kim, J. Park, and J. W. Ha, "Music emotion recognition via end-to-end multimodal neural networks," in *Proc. RecSys*, 2017, p. 11.

[38] J. Dutta and D. Chanda, "Music emotion recognition in assamese songs using MFCC features and MLP classifier," in *Proc. Int. Conf. Intell. Technol. (CONIT)*, Jun. 2021, pp. 1–5.

[39] S. Hizlisoy, S. Yildirim, and Z. Tufekci, "Music emotion recognition using convolutional long short term memory deep neural networks," *Eng. Sci. Technol., Int. J.*, vol. 24, no. 3, pp. 760–767, Jun. 2021.

[40] S. Ma and R. Zhou, "Violin music emotion recognition with fusion of CNN–BiGRU and attention mechanism," *Information*, vol. 15, no. 4, p. 224, Apr. 2024.

[41] H.-T. Hung, J. Ching, S. Doh, N. Kim, J. Nam, and Y.-H. Yang, "EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation," 2021, *arXiv:2108.01374*.

[42] J. Grekow, "Music emotion recognition using recurrent neural networks and pretrained models," *J. Intell. Inf. Syst.*, vol. 57, no. 3, pp. 531–546, Dec. 2021.

[43] J. S. Gómez-Cañón, N. Gutiérrez-Páez, L. Porcaro, A. Porter, E. Cano, P. Herrera-Boyer, A. Gkiokas, P. Santos, D. Hernández-Leo, C. Karreman, and E. Gómez, "TROMPA-MER: An open dataset for personalized music emotion recognition," *J. Intell. Inf. Syst.*, vol. 60, no. 2, pp. 549–570, Apr. 2023.

[44] R. Panda, R. Malheiro, and R. P. Paiva, "Musical texture and expressivity features for music emotion recognition," in *Proc. 19th Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, 2018, pp. 383–391.

[45] S. Gupta, "Deep audio embeddings and attention based music emotion recognition," in *Proc. 15th Int. Conf. Develop. eSystems Eng. (DeSE)*, Jan. 2023, pp. 357–362.

[46] E. Koh and S. Dubnov, "Comparison and analysis of deep audio embeddings for music emotion recognition," 2021, *arXiv:2104.06517*.

[47] R. Malheiro, R. Panda, P. J. S. Gomes, and R. P. Paiva, "Bi-modal music emotion recognition: Novel lyrical features and dataset," in *Proc. 9th Int. Workshop Music Mach. Learn. Eur. Conf. Mach. Learn. Princ. Pract. Knowl. Discovery Databases (ECML/PKDD)*, 2016, pp. 153–160.

[48] L. Xu, Z. Sun, X. Wen, Z. Huang, C.-J. Chao, and L. Xu, "Using machine learning analysis to interpret the relationship between music emotion and lyric features," *PeerJ Comput. Sci.*, vol. 7, p. e785, Nov. 2021.

[49] L. Xu, Z. Yun, Z. Sun, X. Wen, X. Qin, and X. Qian, "PSIC3839: Predicting the overall emotion and depth of entire songs," in *Design Studies and Intelligence Engineering*. Amsterdam, The Netherlands: IOS Press, 2022, pp. 1–9.

[50] D. Chaudhary, N. P. Singh, and S. Singh, "Development of music emotion classification system using convolution neural network," *Int. J. Speech Technol.*, vol. 24, no. 3, pp. 571–580, Sep. 2021.

[51] H. T. P. Thao, G. Roig, and D. Herremans, "EmoMV: Affective music-video correspondence learning datasets for classification and retrieval," *Inf. Fusion*, vol. 91, pp. 64–79, Mar. 2023.

[52] E. Y. Koh, K. W. Cheuk, K. Y. Heung, K. R. Agres, and D. Herremans, "MERP: A music dataset with emotion ratings and raters' profile information," *Sensors*, vol. 23, no. 1, p. 382, Dec. 2022.

[53] K. Hevner, "Experimental studies of the elements of expression in music," *Amer. J. Psychol.*, vol. 48, no. 2, p. 246, Apr. 1936.

[54] M. Chełkowska-Zacharewicz and M. Janowski, "Polish adaptation of the Geneva emotional music scale: Factor structure and reliability," *Psychol. Music*, vol. 49, no. 5, pp. 1117–1131, Sep. 2021.
[55] P. Ekman, "The argument and evidence about universals in facial expressions," in *Handbook of Social Psychophysiology*, vol. 143. New York, NY, USA: Oxford Univ. Press, 1989, p. 164.
[56] I. J. Roseman, M. S. Spindel, and P. E. Jose, "Appraisals of emotion-eliciting events: Testing a theory of discrete emotions," *J. Personality Social Psychol.*, vol. 59, no. 5, pp. 899–915, Nov. 1990.
[57] Y. Feng, Y. Zhuang, and Y. Pan, "Popular music retrieval by detecting mood," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2003, pp. 375–376.
[58] L. Lu, D. Liu, and H.-J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Trans. Audio, Speech Language Process.*, vol. 14, no. 1, pp. 5–18, Jan. 2006.
[59] M. Bilal Er and I. B. Aydilek, "Music emotion recognition by using chroma spectrogram and deep visual features," *Int. J. Comput. Intell. Syst.*, vol. 12, no. 2, p. 1622, 2019.
[60] F. Zhang, H. Meng, and M. Li, "Emotion extraction and recognition from music," in *Proc. 12th Int. Conf. Natural Comput., Fuzzy Syst. Knowl. Discovery (ICNC-FSKD)*, Aug. 2016, pp. 1728–1733.
[61] X. Jia, "Music emotion classification method based on deep learning and improved attention mechanism," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–8, Jun. 2022.
[62] P. R. Farnsworth, "A study of the hevner adjective list," *J. Aesthetics Art Criticism*, vol. 13, no. 1, pp. 97–103, Sep. 1954.
[63] P. R. Farnsworth, *The Social Psychology of Music*. New York, NY, USA: Oxford Univ. Press, 1958.
[64] E. Schubert, "Update of the hevner adjective checklist," *Perceptual Motor Skills*, vol. 96, no. 4, p. 1117, 2003.
[65] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, p. 1161, 1980.
[66] W. Chen, "A novel long short-term memory network model for multimodal music emotion analysis in affective computing," *J. Appl. Sci. Eng.*, vol. 26, no. 3, pp. 367–376, 2022.
[67] B. G. Patra, P. Maitra, D. Das, and S. Bandyopadhyay, "MediaEval 2015: Music emotion recognition based on feed-forward neural network," in *Proc. MediaEval*, 2015.
[68] R. Panda, R. Malheiro, and R. P. Paiva, "Novel audio features for music emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 11, no. 4, pp. 614–626, Oct. 2020.
[69] R. Panda, B. Rocha, and R. P. Paiva, "Dimensional music emotion recognition: Combining standard and melodic audio features," in *Proc. 10th Int. Symp. Comput. Music Multidisciplinary Res.*, 2013, pp. 583–593.
[70] K. Markov and T. Matsui, "Music genre and emotion recognition using Gaussian processes," *IEEE Access*, vol. 2, pp. 688–697, 2014.
[71] Y.-H. Chin, C.-H. Lin, E. Siahaan, I.-C. Wang, and J.-C. Wang, "Music emotion classification using double-layer support vector machines," in *Proc. 1st Int. Conf. Orange Technol. (ICOT)*, Mar. 2013, pp. 193–196.
[72] J. Grekow, "Static music emotion recognition using recurrent neural networks," in *Proc. Int. Symp. Methodologies Intell. Syst.*, Graz, Austria. Cham, Switzerland: Springer, 2020, pp. 150–160.
[73] C. C. Liu, Y. H. Yang, P. H. Wu, and H. Chen, "Detecting and classifying emotion in popular music," in *Proc. 9th Joint Int. Conf. Inf. Sci.*, 2006, pp. 996–999.
[74] Y. Chin, Y. Hsieh, M. Su, S. Lee, M. Chen, and J. Wang, "Music emotion recognition using PSO-based fuzzy hyper-rectangular composite neural networks," *IET Signal Process.*, vol. 11, no. 7, pp. 884–891, Sep. 2017.
[75] M. Malik, S. Adavanne, K. Drossos, T. Virtanen, D. Ticha, and R. Jarina, "Stacked convolutional and recurrent neural networks for music emotion recognition," 2017, *arXiv:1706.02292*.
[76] A. Ualibekova and P. Shamoi, "Music emotion recognition using K-nearest neighbors algorithm," in *Proc. Int. Conf. Smart Inf. Syst. Technol. (SIST)*, Apr. 2022, pp. 1–6.
[77] J. H. Juthi, A. Gomes, T. Bhuiyan, and I. Mahmud, "Music emotion recognition with the extraction of audio features using machine learning approaches," in *Proc. ICETIT*. Cham, Switzerland: Springer, 2020, pp. 318–329.
[78] W. C. Chiang, J. S. Wang, and Y. L. Hsu, "A music emotion recognition algorithm with hierarchical SVM based classifiers," in *Proc. Int. Symp. Comput., Consum. Control*, Jun. 2014, pp. 1249–1252.
[79] R. E. Thayer, *The Biopsychology of Mood and Arousal*. London, U.K.: Oxford Univ. Press, 1990.
[80] A. Tellegen, D. Watson, and L. A. Clark, "On the dimensional and hierarchical structure of affect," *Psychol. Sci.*, vol. 10, no. 4, pp. 297–303, Jul. 1999.
[81] A. Mehrabian, *An Approach to Environmental Psychology*. Cambridge, MA, USA: MIT Press, 1974.
[82] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychol.*, vol. 14, no. 4, pp. 261–292, Dec. 1996.
[83] I. Bakker, T. van der Voordt, P. Vink, and J. de Boon, "Pleasure, arousal, dominance: Mehrabian and Russell revisited," *Current Psychol.*, vol. 33, no. 3, pp. 405–421, Sep. 2014.
[84] Z. Xiao, E. Dellandrea, W. Dou, and L. Chen, "What is the best segment duration for music mood analysis?" in *Proc. Int. Workshop Content-Based Multimedia Indexing*, vol. 240, Jun. 2008, pp. 17–24.
[85] M. J. Lucia-Mulas, P. Revuelta-Sanz, B. Ruiz-Mezcua, and I. Gonzalez-Carrasco, "Automatic music emotion classification model for movie soundtrack subtitling based on neuroscientific premises," *Appl. Intell.*, vol. 53, no. 22, pp. 27096–27109, Nov. 2023.
[86] X. Han, F. Chen, and J. Ban, "Music emotion recognition based on a neural network with an inception-GRU residual structure," *Electronics*, vol. 12, no. 4, p. 978, Feb. 2023.
[87] C. Chen and Q. Li, "A multimodal music emotion classification method based on multifeature combined network classifier," *Math. Problems Eng.*, vol. 2020, pp. 1–11, Aug. 2020.
[88] Y. Xia and F. Xu, "Study on music emotion recognition based on the machine learning model clustering algorithm," *Math. Problems Eng.*, vol. 2022, pp. 1–11, Oct. 2022.
[89] X. Jia, "A music emotion classification model based on the improved convolutional neural network," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–11, Feb. 2022.
[90] G. Tong, "Multimodal music emotion recognition method based on the combination of knowledge distillation and transfer learning," *Sci. Program.*, vol. 2022, pp. 1–13, Feb. 2022.
[91] M. Jandaghian, S. Setayeshi, F. Razzazi, and A. Sharifi, "Music emotion recognition based on a modified brain emotional learning model," *Multimedia Tools Appl.*, vol. 82, no. 17, pp. 26037–26061, Jul. 2023.
[92] T. Krols, Y. Nikolova, and N. Oldenburg, "Multi-modality in music: Predicting emotion in music from high-level audio features and lyrics," 2023, *arXiv:2302.13321*.
[93] T. L. Li, A. B. Chan, and A. H. Chun, "Automatic musical pattern feature extraction using convolutional neural network," *Genre*, vol. 10, p. 1, Apr. 2010.
[94] M. Zhang, Y. Zhu, W. Zhang, Y. Zhu, and T. Feng, "Modularized composite attention network for continuous music emotion recognition," *Multimedia Tools Appl.*, vol. 82, no. 5, pp. 7319–7341, Feb. 2023.
[95] N. Niu, "Music emotion recognition model using gated recurrent unit networks and multi-feature extraction," *Mobile Inf. Syst.*, vol. 2022, no. 1, 2022, Art. no. 5732687.
[96] P. Louro, H. Redinho, R. Malheiro, R. P. Paiva, and R. Panda, "Exploring deep learning methodologies for music emotion recognition," in *Proc. Sound Music Comput. Conf. (SMC)*, 2024, pp. 1–8.
[97] M. M. Bradley and P. J. Lang, "Affective norms for English words (ANEW): Instruction manual and affective ratings," Dept. Center Res. Psychophysiology, Univ. Florida, Gainesville, FL, USA, Tech. Rep. C-1, 1999.
[98] M.-H. Su, C.-H. Wu, K.-Y. Huang, and Q.-B. Hong, "LSTM-based text emotion recognition using semantic and emotional word vectors," in *Proc. 1st Asian Conf. Affect. Comput. Intell. Interact. (ACII Asia)*, May 2018, pp. 1–6.
[99] A. S. Sams and A. Zahra, "Multimodal music emotion recognition in Indonesian songs based on CNN-LSTM, XLNet transformers," *Bull. Electr. Eng. Informat.*, vol. 12, no. 1, pp. 355–364, Feb. 2023.
[100] J. W. Pennebaker, "Linguistic inquiry and word count: LIWC 2001," Tech. Rep., 2001. [Online]. Available: http://downloads.liwc.net.s3.amazonaws.com/LIWC2015_OperatorManual.pdf
[101] J. Xu, X. Li, Y. Hao, and G. Yang, "Source separation improves music emotion recognition," in *Proc. Int. Conf. Multimedia Retr.*, Apr. 2014, pp. 423–426.
[102] Y. J. Hsu and C. P. Chen, "Going deep: Improving music emotion recognition with layers of support vector machines," in *Proc. Int. Conf. Appl. Syst. Innov. (ICASI)*, 2015, pp. 209–212.
[103] Y.-Y. Shi, X. Zhu, H.-G. Kim, and K.-W. Eom, "A tempo feature via modulation spectrum analysis and its application to music emotion classification," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2006, pp. 1085–1088.

[104] B. Han, S. Rho, R. B. Dannenberg, and E. Hwang, "SMERS: Music emotion recognition using support vector regression," in *Proc. ISMIR*, 2009, pp. 651–656.
[105] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 2, pp. 448–457, Feb. 2008.
[106] Y. Deng, Y. Lv, M. Liu, and Q. Lu, "A regression approach to categorical music emotion recognition," in *Proc. IEEE Int. Conf. Prog. Informat. Comput. (PIC)*, Dec. 2015, pp. 257–261.
[107] J. Bai, K. Luo, J. Peng, J. Shi, Y. Wu, L. Feng, J. Li, and Y. Wang, "Music emotions recognition by cognitive classification methodologies," in *Proc. IEEE 16th Int. Conf. Cognit. Informat. Cognit. Comput. (ICCI)*, Jul. 2017, pp. 121–129.
[108] S. Fukayama and M. Goto, "Music emotion recognition with adaptive aggregation of Gaussian process regressors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 71–75.
[109] S.-H. Chen, Y.-S. Lee, W.-C. Hsieh, and J.-C. Wang, "Music emotion recognition using deep Gaussian process," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, Dec. 2015, pp. 495–498.
[110] F. H. Rachman, R. Sarno, and C. Fatichah, "Music emotion classification based on lyrics-audio using corpus based emotion," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 3, p. 1720, Jun. 2018.
[111] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[112] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, nos. 5–6, pp. 602–610, Jul. 2005.
[113] M. Velankar, S. Thombre, and H. Wadkar, "Evaluating deep learning models for music emotion recognition," *Int. J. Eng. Appl. Sci. Technol.*, vol. 7, no. 6, pp. 252–259, Oct. 2022.
[114] I.-S. Huang, Y.-H. Lu, M. Shafiq, A. Ali Laghari, and R. Yadav, "A generative adversarial network model based on intelligent data analytics for music emotion recognition under IoT," *Mobile Inf. Syst.*, vol. 2021, pp. 1–8, Nov. 2021.
[115] Y. R. Pandeya and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," *Multimedia Tools Appl.*, vol. 80, no. 2, pp. 2887–2905, Jan. 2021.
[116] G. Liu and Z. Tan, "Research on multi-modal music emotion classification based on audio and lyirc," in *Proc. IEEE 4th Inf. Technol., Netw., Electron. Autom. Control Conf. (ITNEC)*, vol. 1, Jun. 2020, pp. 2331–2335.
[117] S. Yang, D. He, and M. Zhang, "A speaker system based on CLDNN music emotion recognition algorithm," in *Proc. ICETIS ; 7th Int. Conf. Electron. Technol. Inf. Sci.*, Jan. 2022, pp. 1–7.
[118] D. de Matos, W. Ramos, M. Silva, L. Romanhol, and E. R. Nascimento, "A multimodal hyperlapse method based on video and songs' emotion alignment," *Pattern Recognit. Lett.*, vol. 166, pp. 174–181, Feb. 2023.
[119] G. Tong, "Music emotion classification method using improved deep belief network," *Mobile Inf. Syst.*, vol. 2022, pp. 1–7, Mar. 2022.
[120] J. Kang, H. L. Wang, and G. B. Su, "Survey of music emotion recognition," *Comput. Eng. Appl.*, vol. 58, no. 4, pp. 64–72, 2022.

XINGGUO JIANG received the M.S. degree from Chongqing University, Chongqing, China, in 2003, and the Ph.D. degree from the Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu, China, in 2007. From February 2012 to February 2013, he visited Tulane University for an academic exchange. He is currently an Associate Professor with the School of Automation and Electrical Information, Sichuan University of Science and Engineering, Sichuan, China. Prior to that, he was an Associate Professor with the School of Information and Communication, Guilin University of Electronic Technology, Guangxi, China. His current research interests include image processing, intelligent information processing, and deep learning.

YUCHAO ZHANG received the B.S. degree from Sichuan University of Science and Engineering, Sichuan, China, in 2023, where he is currently pursuing the degree with the School of Automation and Electrical Information. His current research interests include natural language processing, music emotion recognition, and deep learning.

GUOJUN LIN received the M.S. degree from Southwest Jiaotong University, Chengdu, China, in 2008, and the Ph.D. degree from the University of Electronic Science and Technology, Chengdu, in 2014. From November 2015 to October 2016, he was a Research Assistant with the School of Computer and Software, Shenzhen University. He is currently a Lecturer with Sichuan University of Science and Engineering. He is mainly working in the direction of face recognition, image processing, and computer vision.

LING YU received the B.S. degree from Sichuan University of Science and Engineering, Sichuan, China, in 2023, where she is currently pursuing the degree with the School of Automation and Electrical Information. Her current research interests include computer vision, pedestrian re-identification, and deep learning.