A Survey on Multimodal Music Emotion Recognition

RASHINI LIYANARACHCHI, University of New South Wales, Australia

ADITYA JOSHI, University of New South Wales, Australia

ERIK MEIJERING, University of New South Wales, Australia

Multimodal music emotion recognition (MMER) is an emerging discipline in music information retrieval that has experienced a surge in interest in recent years. This survey provides a comprehensive overview of the current state-of-the-art in MMER. Discussing the different approaches and techniques used in this field, the paper introduces a four-stage MMER framework, including multimodal data selection, feature extraction, feature processing, and final emotion prediction. The survey further reveals significant advancements in deep learning methods and the increasing importance of feature fusion techniques. Despite these advancements, challenges such as the need for large annotated datasets, datasets with more modalities, and real-time processing capabilities remain. This paper also contributes to the field by identifying critical gaps in current research and suggesting potential directions for future research. The gaps underscore the importance of developing robust, scalable, and interpretable models for MMER, with implications for applications in music recommendation systems, therapeutic tools, and entertainment.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; *Natural language processing*; *Computer vision*; *Machine learning*; • **Applied computing** → **Sound and music computing**.

Additional Key Words and Phrases: Multimodal Music Emotion Recognition, Music Information Retrieval, Deep Learning, Multimodal Feature Extraction, Feature Fusion, Recommendation Systems.

1 Introduction

The emotional impact of music is profound and pervasive, influencing the listener's mood, behavior, and cognitive functions [94]. The ability to automatically detect emotions in music has significant implications across various domains, including music recommendation systems, therapeutic applications, multimedia content creation and analysis, human-computer interaction (HCI), along with other commercial applications. Understanding emotional indicators can enhance user experiences in music streaming services, where personalized recommendations based on emotional preferences can lead to more satisfying and engaging interactions [51]. In therapeutic settings, music is often used as a tool for emotional expression and regulation, making accurate emotion detection crucial for developing effective music-based interventions [54]. Additionally, in multimedia content creation and analysis, identifying the emotional tone of music can aid in selecting the right soundtrack to evoke desired emotions in films, advertisements, and other media projects. This precision ensures that the audio complements the visual narrative, amplifying the overall, desired emotional impact [101]. In HCI, understanding the emotions conveyed by music can enhance the emotional responsiveness and empathy of systems, fostering more immersive and personalized user experiences. By interpreting the emotional intent of music,

Authors' Contact Information: Rashini Liyanarachchi, University of New South Wales, Sydney, Australia, r.liyanarachchi_lekamlage@unsw.edu.au; Aditya Joshi, University of New South Wales, Sydney, Australia, aditya.joshi@unsw.edu.au; Erik Meijering, University of New South Wales, Sydney, Australia, erik.meijering@unsw.edu.au.

systems can dynamically adjust to the user's emotional state, offering mood-aligned content, improving engagement, and delivering more supportive experience [107]. Moreover, in commercial applications, brands can leverage emotion recognition to align their marketing strategies with the consumer's emotional state, enhancing brand resonance and loyalty while fostering more personalized and impactful consumer relationships [22]. Artificial intelligence (AI)-based music creation and synthesis further expand creative possibilities by generating original compositions [53] and adapting music to specific emotional contexts [38], providing a powerful tool for both artistic expression and commercial use. This relies on the ability to detect emotion in music.

Music emotion recognition (MER) is a multidisciplinary field at the intersection of musicology, psychology, and computer science, aiming to identify and classify emotions expressed in music. From a technical standpoint, MER involves the analysis of musical elements such as melody, harmony, rhythm, and lyrics. Traditional methods for MER focus primarily on audio features extracted from the audio signal, but recent advancements emphasize the importance of multimodal approaches, incorporating additional modalities such as lyrics, visuals, user context, and physiological responses [60, 73, 90, 105, 117, 118]. In recent years, research in MER has witnessed a rapid growth, with the advancement of AI and the increased accessibility of digital music. Multiple surveys have been conducted on MER [30, 42, 48, 55, 107], focusing primarily on audio features and traditional machine learning algorithms [48, 55, 107] or, more recently, deep learning [30]. Systematic evaluations on automated MER as a continuous regression problem in the arousal-valence (AV) plane using audio have also been presented [42]. However, past literature lacks an overview of the state-of-the-art in MER using not only audio but also other modalities.

Our paper aims to fill the current literature gap by providing a comprehensive survey on approaches to multimodal MER (MMER). These approaches integrate multiple modalities such as audio, lyrics, video, physiological signals, and others, to achieve more accurate and reliable results. Additionally, being a fast-evolving field, this paper delves into more recent advancements in deep learning-based MMER since 2022, highlighting the latest techniques that leverage multimodal data for more accurate and reliable MER. We start by summarizing the necessary background and preliminaries relevant to understand what MER is about (Section 2). Then we discuss the methods and techniques in the process of MMER in terms of the previous literature (Section 3). Next, we highlight current trends in MMER, challenges, and future directions (Section 4). Finally, we summarize the main conclusions and take-home lessons (Section 5).

2 Background and Preliminaries

Before delving deeper into MMER methods and challenges, we first discuss different theories of emotion in music (Section 2.1), as well as important emotion models (Section 2.2), the primary modalities used in MMER (Section 2.3), publicly available datasets for music emotion recognition (Section 2.4), and metrics for method evaluation (Section 2.5).

2.1 Theories of Emotion and Music

To design and develop effective computational models for MER, it is crucial to comprehend the connection between emotion and music. Numerous features of music have been identified as indicative of distinct emotions. Basic emotions (Table 1) can be conveyed by specific musical features, such as [10]:

- Tempo: The speed at which a piece of music is played.
- Mode: The scale or tonal framework used in a piece such as major or minor.
- Harmony: The simultaneous perception of two or more notes producing a pleasing sound.
- Sound Level: The amplitude or intensity at which the musical composition is performed.

Emotion	Musical Features
Happiness	Fast tempo, Small tempo variability, Major mode, Simple and consonant harmony, Medium-high sound level, Small sound level variability, High pitch, Ascending pitch, Perfect 4th and 5th intervals, Staccato articulation
Sadness	Slow tempo, Minor mode, Dissonance, Low sound level, Moderate sound level variability, Low pitch, Descending pitch, Small intervals (e.g., minor 2nd), Legato articulation
Anger	Fast tempo, Small tempo variability, Minor mode, Dissonance, High sound level, Small loudness variability, High pitch, Ascending pitch, Major 7th and augmented 4th intervals, Staccato articulation
Fear	Fast tempo, Large tempo variability, Minor mode, Dissonance, Low sound level, Large sound level variability, High pitch, Ascending pitch, Staccato articulation
Tenderness	Slow tempo, Major mode, Medium-low sound level, Small sound level variability, Low pitch, Legato articulation
Surprise	Fast tempo, Large tempo variability, Major mode, High sound level, Large sound level variability, High pitch, Sudden dynamic changes, Staccato articulation, Unexpected pauses or rhythmic changes

Table 1. Summary of musical features correlated with discrete emotions. Features for all emotions except "Surprise" are adopted from Patrik et al. [50].

- Pitch: The position of an individual sound within the entire spectrum of sound.
- Interval: The discrete change from one pitch to another (major 3rd, perfect 5th, etc. [93]).
- Articulation: The way notes are played in sequence (legato, staccato, etc. [11]).
- Rhythm: The pattern of beats and how they are grouped together.
- Melody: The sequence of notes that are perceived as a single entity.
- Dynamics: The volume of the music, including changes in loudness and softness.

However, the same feature can be used in a similar manner to express different emotions. As an example, a fast tempo can be used to express happiness, anger, or fear. Thus, each feature in and of itself is neither necessary nor conclusive. Despite this ambiguity, the greater the number of indicators employed, the more dependable the communication becomes [49].

2.2 Emotion Models

Understanding and utilizing emotion models is fundamental in MER. Emotion models provide frameworks for categorizing and interpreting the complex and often subjective nature of emotions conveyed through music, facilitating more consistent and comparable results across research and applications. Among the widely recognized models (Table 2), Russell's Circumplex Model [79] and Thayer's Model [92] are noteworthy here. Russell's model maps emotions onto a circular structure defined by valence (expressing feelings from negative to positive) and arousal (expressing the intensity of the emotion from low to high), offering a continuous and dynamic perspective on emotional experiences (Fig. 1). Thayer's Model, in contrast, introduces the dimensions of energy-stress and calm-tired, focusing on the physiological aspects of emotions.

Other models include Plutchik's Wheel of Emotions [75] and Ekman's Basic Emotions [25]. Plutchik's model arranges eight primary emotions in a wheel, highlighting their combinations and interactions, while Ekman's model identifies six universal emotions: happiness, sadness, fear, disgust, anger, and surprise. These discrete emotion models provide a straightforward approach to categorizing emotional content. More complex models such as Scherer's Component Process Model (CPM) [82] and Lindquist's Conceptual Act Model [62] emphasize the dynamic and constructed nature of emotions. CPM focuses on the dynamic changes in emotion components, whereas Lindquist's model integrates core affect with conceptual knowledge, underscoring the role of cognitive processes in emotional experiences.

Additionally, models such as Positive Activation Negative Activation (PANA) [102] and Geneva Emotion Wheel (GEW) [83] offer nuanced insights into how emotions are constructed and experienced. Barrett's theory [6] suggests that emotions are constructed in the moment by core affect and conceptual knowledge, while the GEW maps complex emotions in a circular manner, reflecting their multidimensional nature. The PANA model distinguishes between positive and negative activation dimensions.

Furthermore, Lazarus' Cognitive-Mediational Theory [61] highlights the critical role of cognitive appraisal in emotional responses. According to this theory, emotions arise from an individual's evaluation of an event's significance

Model	Classes / Dimensions	Categorical / Dimensional	Domain
Russell's Circumplex Model [79]	2	Dimensional	General
Thayer's Model [92]	2	Dimensional	General
Plutchik's Wheel of Emotions [75]	8	Categorical	General
Ekman's Basic Emotions [25]	6	Categorical	General
Scherer's Component Process Model [82]	Varies	Dimensional	General
Lindquist's Conceptual Act Model [62]	Varies	Both	General
Positive Activation Negative Activation Model [102]	2	Dimensional	General
Geneva Emotion Wheel [83]	20	Categorical	General
Barrett's Model [6]	Varies	Dimensional	General
Lazarus's Cognitive-Mediational Theory [61]	Varies	Both	General
Watson and Tellegen's Circumplex Model of Affect [102]	2	Dimensional	General
Geneva Emotional Music Scale Model [110]	9	Categorical	Music
Hevner's Emotional Model [76]	8	Categorical	Music

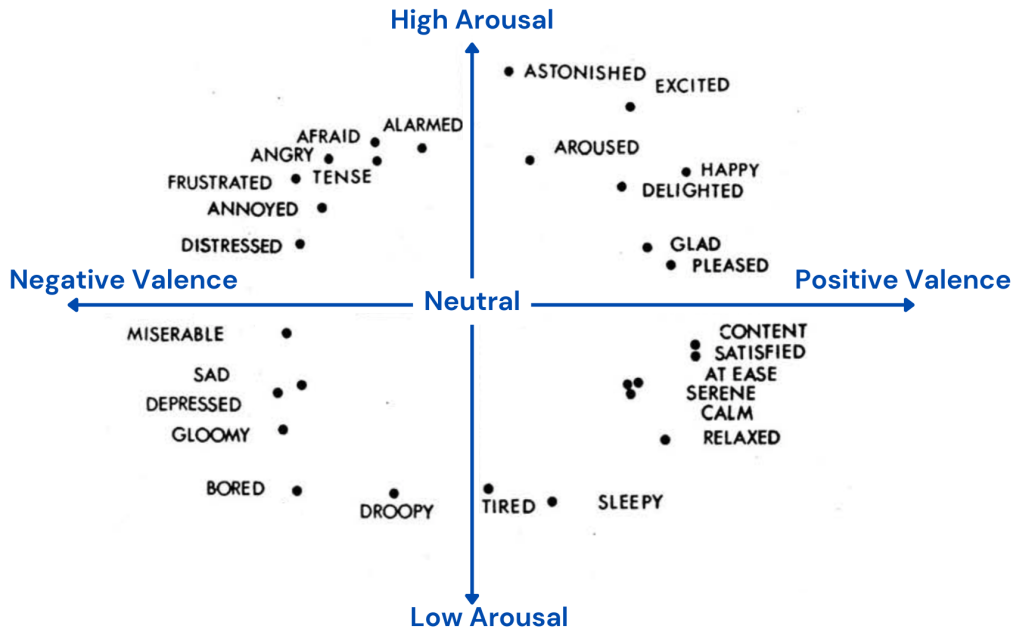Table 2. Emotion models used in MER.

Fig. 1. Russell's Circumplex Model. Emotions are characterized by valence (ranging from negative to positive) along the horizontal axis and arousal (ranging from low to high) along the vertical axis.

and their ability to cope with it. This appraisal process determines the emotional reaction and subsequent coping strategies, providing a framework for understanding how personal interpretation influences emotional experiences.

Moreover, Watson and Tellegen's Circumplex Model of Affect [102], similar to Russell's model, emphasizes positive and negative affects. Each of these models contributes unique perspectives, enhancing our understanding of the intricate relationship between music and emotions.

Two emotion models have been used in the context of music, namely, the Geneva Emotional Music Scale (GEMS) [110] and Hevner's Emotional Model [76]. The GEMS model identifies distinct emotional responses such as wonder, transcendence, tenderness, nostalgia, peacefulness, power, joyful activation, tension, and sadness, providing a tailored framework for exploring the rich and nuanced emotional landscape evoked by musical experiences. In contrast, Hevner's model classifies music emotions based on a set of adjectives divided into eight groups. Although this discrete model is less smooth than Thayer's, it is useful in distinguishing between specific emotional states described in musical terms.

2.3   Modalities Used in MMER

MMER is a multidisciplinary field that uses various modalities to understand and classify the emotional content of music. By combining different sources of information, researchers can develop more accurate and nuanced models. The primary modalities used in MMER are (Fig. 2):

(1) **Audio:** This refers to the sonic aspect of music, including pitch, loudness, and tempo. It is used to understand and identify the emotions conveyed through the musical piece.
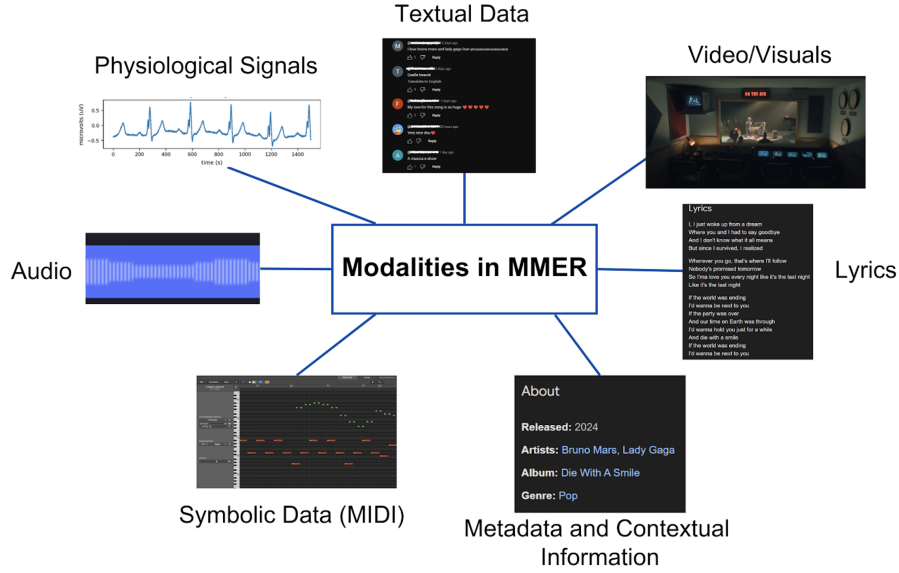
Fig. 2. Modalities used in MMER.

(2) **Lyrics:** Textual content of songs play a role in conveying underlying emotions. Features such as word choice, themes, and sentiment can be used to identify and interpret the emotions conveyed by the lyrics.

(3) **Visuals:** Music is often accompanied by videos, live performances, or animations containing visual indicators like color, lighting, facial expressions, and body language, which contribute to the overall emotional experience of the music.

(4) **Symbolic Data:** MIDI (Musical Instrument Digital Interface) is a mode of music representation that uses discrete events and parameters such as notes, dynamics, and tempo. It enables studying musical structure and patterns to determine emotional content. This symbolic data can be mapped to acoustic features, such as sound intensity, pitch variation, and rhythmic patterns, providing a bridge between the abstract representation of music and its actual auditory characteristics, thereby enhancing the analysis of emotional expression.

(5) **Physiological Signals:** Emotional responses to music can be measured in the form of changes in skin conductance and heart rate or by electroencephalography (EEG) and electrocardiography (ECG). These measurements provide quantitative data about the physiological effects of music on the listener's qualitative emotional state.

(6) **Textual Data:** Apart from lyrics, other textual data such as reviews, social media posts, and comments about songs can be used to analyze and understand the emotional impact of music. They capture subjective emotional responses and sentiments from listeners.

(7) **Metadata and Contextual Data:** Further information about music, such as genre, artist, release year, and cultural context, is often provided in the metadata of digital songs. The background and contextual understanding to which this information contributes enhances the emotional interpretation of a song.

Dataset	Year	Size	Annotation	Modalities	Labeling
CAL500 [96]	2008	500	SLEA	Audio	Categorical
Soundtracks [24]	2011	470	SLEA	Audio	Both
DEAP [57]	2012	120	SLEA	Video, EEG, Physiological	Dimensional
MediaEval Emotion in Music [85]	2013	744	Both	Audio	Dimensional
MIREX-like [72]	2013	193	SLEA	Audio, Lyrics, MIDI	Categorical
CAL500exp [98]	2014	3,223	CEA	Audio	Categorical
AMG1608 [15]	2015	1,608	SLEA	Audio	Dimensional
NJU [104]	2015	777	SLEA	Audio, Lyrics	Categorical
Emotify [2]	2016	400	SLEA	Audio	Categorical
Audioset [27]	2017	16,955	SLEA	Video	Categorical
DEAM [3]	2017	1,802	Both	Audio, Metadata	Dimensional
4Q Dataset [71]	2018	900	SLEA	Audio	Categorical
PMEmo [111]	2018	794	Both	Audio, Lyrics, Comments, Physiological	Dimensional
The MTG-Jamendo [9]	2019	55,525	SLEA	Audio	Categorical
Music4all [74]	2020	109,269	Both	Audio	Dimensional
EMOPIA [41]	2021	1,087	SLEA	MIDI, Metadata	Categorical
MuSe [1]	2021	90,001	SLEA	Audio	Dimensional
HKU956 [36]	2022	956	SLEA	Audio, Physiological signals	Dimensional
MusAV [8]	2022	2,092	SLEA	Audio	Dimensional
MuVi [17]	2022	81	CEA	Audio, Video (Muted)	Dimensional
TROMPA-MER [28]	2022	1,161	SLEA	Audio	Categorical
emoMV [91]	2023	5,986	SLEA	Audio, Video	Categorical
MERP [58]	2023	54	CEA	Audio, Metadata	Dimensional
EMMA [86]	2024	364	SLEA	Audio, Video	Categorical
SiTunes [29]	2024	300	SLEA	Audio, Physiological, Weather	Dimensional
MERGE [66]	2024	2,000	SLEA	Audio, Lyrics	Dimensional

Table 3. Publicly available MER datasets. Size indicates the number of songs or fragments in the dataset. Annotation includes SLEA (song-level emotion annotation), CEA (continuous emotion annotation), or both. Labeling includes categorical (emotions classified into distinct categories such as happy or sad), dimensional (emotions represented on a continuous scale such as valence or arousal), or both.

2.4 Music Emotion Recognition Datasets

Many datasets are available for MER research (Table 3). However, most of them are unimodal, typically containing only audio. Broadly speaking, existing datasets can be divided into two categories: song-level emotion annotation (SLEA) and continuous emotion annotation (CEA). The annotation labels used in each case can be categorical (distinct emotions) or dimensional (numerical values over a number of emotion dimensions such as arousal and valence [79]). The latter is particularly useful for capturing the nuanced and dynamic nature of musical emotions.

2.4.1 Song-Level Emotion Annotation (SLEA). Most datasets use song-level annotation, where a single emotion is assigned to a complete song. These datasets typically contain annotations that label each song with the predominant emotion. The assumption that a song contains a single emotion may not always be true but may work well as a simplification when using MER in a downstream task such as music information retrieval or recommendation.

2.4.2 Continuous Emotion Annotation (CEA). In contrast to song-level annotation, continuous annotation provides time-varying labels that capture the dynamic nature of emotions throughout a song. The time intervals for these annotations are typically determined based on the granularity required for accurate emotion tracking, often ranging from 1 to 5 seconds. Shorter intervals provide a finer resolution of emotional changes but require more detailed analysis,

while longer intervals may capture broader emotional trends but with less temporal precision. The choice of interval depends on the specific objectives of the study and the nature of the music content being analyzed. These datasets are crucial for studies on music emotion variation detection (MEVD).

2.5   Evaluation Metrics

Evaluation metrics are essential for assessing the performance and robustness of MMER methods. MMER may be modeled as a classification or a regression task. For the classification formulation of MMER, Accuracy (A) is a commonly used metric representing the probability of correct classification within a dataset, and provides a basic measure of overall correctness. It is calculated as the ratio of correctly predicted outcomes to the total number of predictions made. Precision (P) and recall (R) are useful for evaluating the reliability and completeness of positive predictions, respectively. Precision measures the ratio of correctly predicted positive observations to the total predicted positives, while recall measures the ratio of correctly predicted positives to all actual positives. The F1 score, being the harmonic mean of precision and recall, is particularly useful for handling imbalanced datasets, offering a balanced view of the model's performance [34, 84].

For regression tasks in MMER, metrics such as the mean absolute error (MAE), the coefficient of determination (R^2), and the root mean squared error (RMSE) are frequently used. MAE estimates the average magnitude of prediction errors, allowing a direct interpretation of the average error. R^2 evaluates the degree to which a given regression model accurately represents the sample data. RMSE, on the other hand, prioritizes greater errors, making it more vulnerable to outliers. The concordance correlation coefficient (CCC) is also used to assess the agreement between predicted and actual values by integrating precision and accuracy into a single metric. Additionally, area under the receiver operating characteristic curve (AUROC) is employed to evaluate the classifier's ability to distinguish between classes at various threshold settings, providing a comprehensive measure of separability. In ranking tasks, metrics such as hits score on top k (hits@k) and mean average precision on top k (map@k) are utilized to evaluate the accuracy and order of a model's top predictions, respectively. These metrics collectively provide a robust framework for evaluating the diverse aspects of MMER performance [30, 88].

Furthermore, public competitions and challenges provide researchers in the field with benchmarks to evaluate and compare methods. These challenges utilize standardized datasets, consistent evaluation protocols and metrics, and allow competitors to rank their methods against the state of the art. Currently, there are no established benchmarks developed specifically for MMER. The only available benchmarks are for audio-based MER. Some of these include the "Emotion and Themes in Music" task in MediaEval[1] (most recent: 2021) and the "Audio K-POP Mood Classification" task in MIREX[2] (most recent: 2020). However, both are based on audio classification only.

Selecting the appropriate evaluation metrics is very important for effective performance assessment and improvement. The chosen metrics should align with the task-specific characteristics and the nature of the dataset. To obtain a comprehensive understanding of the strengths and weaknesses of their models, researchers can utilize a combination of the discussed metrics.

3   Methods and Techniques

MMER research can be summarized using a four-stage framework (Fig. 3), proceeding from modality and data selection (Stage 1), to feature extraction (Stage 2) and feature processing (Stage 3), and, finally, emotion prediction (Stage 4).

---

[1]https://multimediaeval.github.io/editions/2021/tasks/music/ (Accessed 15 November 2024).
[2]https://www.music-ir.org/mirex/wiki/2020:Audio_K-POP_Mood_Classification (Accessed 15 November 2024).
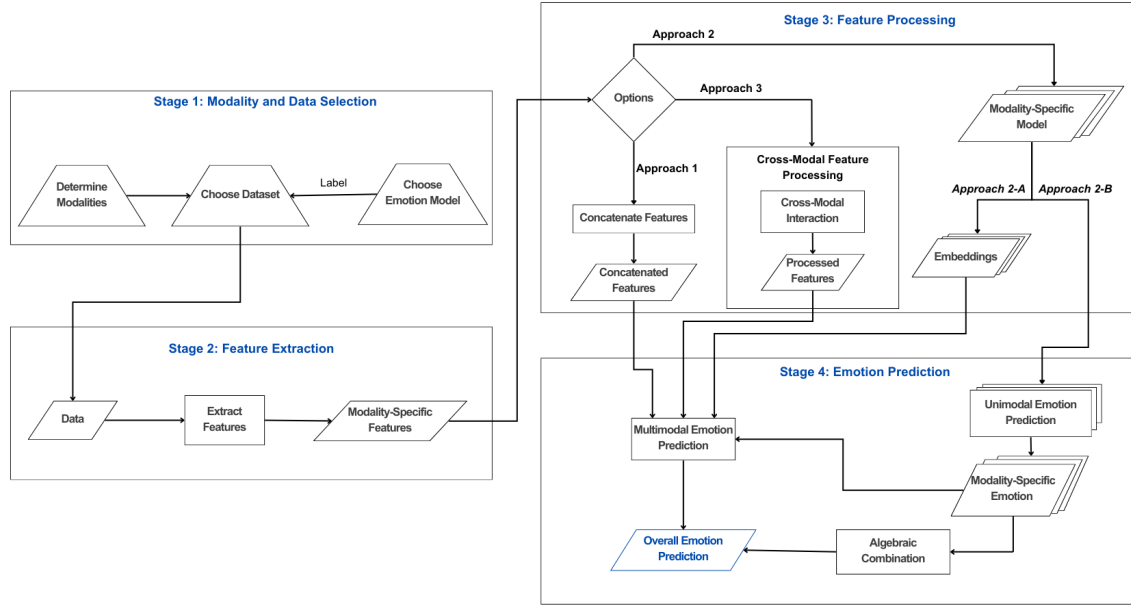
Fig. 3. Framework summarizing past and current MMER methods.

While Stages 1 and 2 exhibit significant similarities across all published works, Stage 3 typically uses one of three distinct approaches, while Stage 4 commonly involves one of two distinct approaches. Having already introduced the various aspects of Stage 1 (especially in Sections 2.2–2.4), we now proceed to discuss methods and techniques involved in Stage 2 (Section 3.1), Stage 3 (Section 3.2), and Stage 4 (Section 3.3).

3.1 Feature Extraction

Feature extraction is a crucial step in MMER, as it transforms data from different modalities into representations of relevant information about a given piece of music that can be used as a basis for further processing and ultimate prediction. In this section we present a detailed overview of feature extraction in MMER, including the data formats, methods and techniques used to obtain audio features, lyric features, visual features, symbolic features, physiological features, textual (user-generated content) features, and metadata/contextual features.

3.1.1 Audio Features. These features are the most widely studied in MER as well as MMER. The key audio elements conveying or expressing emotion in music are tempo, rhythm, mode, melody, and dynamics. Fu et al. [26] divide audio features into three categories: low-level, mid-level, and top-level (track-level) features/labels (Fig. 4). Semantic labels, at the top-level, offer insights into the ways in which people perceive and interpret music like style, genre, emotion, etc. As obtaining these levels is difficult from lower-level features, it makes the top-level features abstract [115]. Therefore, in this survey, we primarily focus on low- and mid-level features, as these serve as the foundation for extracting top-level features like emotion. Top-level features and fusion are included in the discussion of emotion prediction (Section 3.3), aligning with the ultimate objective of understanding emotional responses to music.

Low-Level Features. These can be broadly categorized as spectral features (SFs) and temporal features (TFs). The spectral characteristics of music include the timbre and tonal properties and are captured by SFs present in a relatively short time interval [46]. SFs are obtained using signal processing techniques such as Fourier transformation, spectral/cepstral analysis, etc. Furthermore, the SFs commonly used in MER consist of two types: spectrograms and low-level descriptors (LLDs). The short-time Fourier transform (STFT) is used to generate spectrograms, and the input for models using these is in the form of a two-dimensional (2D) map consisting of a sequence of short-time spectrograms. LLDs calculated using single-frame audio are further processed to obtain high-level statistical functions (HSFs) [13]. Examples of LLDs are the mel-frequency cepstrum coefficient (MFCC) [65], zero crossing rate (ZCR), spectral centroid, spread, roll-off, and flux, and chroma features [47]. Examples of HSFs are the maximum, mean, and variance. In recent literature, the most commonly used timbre features are MFCCs, ZCR, and chroma features, while the centroid is the most widely used spectral feature (Table 4). However, some authors have used unique features, such as joined filter banks (an intermediate product between mel-spectrograms and MFCCs) and 60 handcrafted features to generate input for the classification model [113].

TFs, on the other hand, capture the time-related aspects of audio signals, such as rhythm, tempo, and energy dynamics that are essential for expressing emotions in music. These features capture aspects such as beat consistency, tempo variations, and loudness fluctuations. Examples of TFs include short-time energy (STE), tempo, and beat strength, which characterize the rhythmic and dynamic elements of a musical piece. Unlike SFs, TFs focus on how these audio characteristics evolve over time. For example, the progression of energy or the consistency of rhythmic patterns can significantly affect how arousal or emotional intensity is perceived [107].

The integration of TFs with SFs enhances the representation of musical content by combining instantaneous timbral information with temporal progression, leading to more effective emotion recognition systems. Whether considering SFs or TFs, advanced methods such as convolutional neural networks (CNNs) and recurrent neural network (RNNs) automatically learn hierarchical feature representations, capturing complex patterns and temporal dependencies at multiple scales that are indicative of emotions [16].

Mid-Level Features. In MMER, the use of mid-level features is not as common as in other areas such as cover-song detection. The three most used mid-level features are rhythm, pitch, and harmony. Rhythm is described using two important indicators: beat and tempo (beats-per-minute). The tempo can correspond with other features such as pitch, which is also a mid-level feature, in order to recognize an emotion (see Table 1).

3.1.2 Lyric Features. The lyrics of a song play a crucial part in expressing its emotional charge. Most research uses natural language processing (NLP) models as classifiers or regressors. Therefore, the words of the lyrics need to be converted into specific features to be fed as input to these models. Pyrovolakis et al. [77] used several methodologies such as bag of words (BoW), term frequency-inverse document frequency (TF-IDF), Word2Vec, GloVe, and bidirectional encoder representations from transformers (BERT) embeddings. While BoW is one of the most widely used features in MER, part-of-speech (PoS) tags are also used as features that attribute the grammatical class to each word. However, according to Malheiro et al. [67], BoW and PoS are not sufficient for MER. Therefore, they proposed three new features: presence of slang or colloquial words, structural analysis features, and semantic features. The sequence of words (unigram, bigram, and trigram) is another type of lyric features used in MER research [19]. Chen and Li [13] used two main features: word embeddings and word frequency vectors using Word2Vec and chi-squared test feature extraction respectively. Similarly, Zhang and Tian [112] used chi-squared test feature extraction along with TF-IDF to extract the lyric features to train their models. To enhance the previously mentioned techniques, particularly for handling
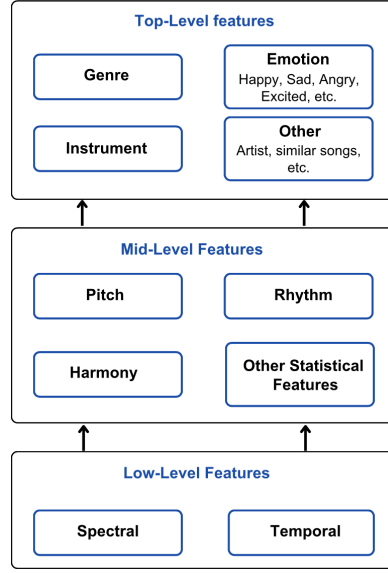
Manuscript

Fig. 4. Categorization of audio features.

Reference	Feature		Year
	Timbre	**Spectral**	
Chen and Li [13]	MFCC, ZCR, Chroma	Spectral Centroid, Spectral Spread, Spectral Roll-Off, Spectral Flux	2020
Liu and Tan [63]	MFCC	Spectrum Centroid	2020
Pandeya et al. [73]	MFCC, ZCR	-	2021
Chen [14]	MFCC	-	2022
Liu et al. [64]	MFCC, ZCR, Chroma	Spectral Centroid, Spectral Roll-Off, Spectral Flatness, Spectral Contrast	2022
Pyrovolakis et al. [77]	MFCC, Chroma	Spectral Contrast	2022
Zhang and Tian [112]	MFCC, ZCR, Chroma	Spectral Centroid, Spectral Bandwidth	2022
Zhang et al. [113]	Filter Bank	-	2022
Zhao and Yoshii [118]	MFCC	-	2023
Wang et al. [97]	MFCC	Spectral Centroid	2024

Table 4. Summary of audio features used in recent literature (2020-2024).

multi-emotion classification problems, Edmonds and Sedoc [23] utilized the NRC Hashtag Emotion Lexion [68] to convert lyrical data into feature vectors with a length of 9. Wang et al. [99] incorporated rhyme as a feature using a rhyme system they created for their study on emotion detection in Chinese songs.

3.1.3 Visual Features. In recent developments within MMER, researchers have begun incorporating visuals, particularly music videos, into their analysis. This addition allows for a richer and more nuanced understanding of the emotional

content of music by combining auditory and visual cues, thereby enhancing the accuracy of emotion recognition systems. Visual features can be divided into two main categories based on the input type: static and dynamic. Appearance-based feature extraction approaches are commonly used for static images. Gabor wavelet representation and local binary patterns (LBP) are two of these methods. Recently, researchers have used CNNs to extract facial features from static images [21, 100, 114]. With regard to dynamic features, Pandeya et al. [73] proposed a method for analyzing video segments that contain only faces. They extracted these using the cascade classifier feature of OpenCV[3]. Facial expression is a visual feature used in the majority of MER literature. Chen [14] used the 68 face feature points obtained using the Dlib[4] Python library to extract these facial expressions. A common strategy to capture the dynamics of movement within a video is to employ optical flow and motion analysis [69]. To infer emotions more accurately, researchers have used context-aware networks that consider the scene's broader context, including the environment and character interactions [73]. Additionally, slow-fast networks, which process video data at multiple temporal resolutions, capture both detailed spatial information and rapid temporal changes [64].

3.1.4 Symbolic Features. These features are extracted from symbolic music scores such as MIDI. Most symbolic music scores used in MER are represented in MIDI, which is an informative resource that provides note information (pitch, duration, etc.), timing information (tempo and time signature), and other relevant details. However, in the case of multimodal or unimodal MER, MIDI files are subjected to additional processing in order to extract symbolic domain features such as melody, rhythm, and other dynamic information. Zhao and Yoshii [118] utilized a RNN, specifically a bidirectional gated recurrent unit (BiGRU), to extract symbolic domain features from MIDI in order to pass them to the emotion classifier. Others, such as Thammasan [90], employed the MIRToolbox developed by Laurier et al. [59] to extract musical features from MIDI files.

3.1.5 Physiological Features. These can be derived from several sources, including EEG, ECG, heart rate variability (HRV), galvanic skin response (GSR), electromyography (EMG), and respiratory rate (RSP) [12] in both multimodal and unimodal MER.

EEG, in particular, is widely used in the field of MER [18] as it captures the brain's electrical activity and provides insights into the emotional responses elicited by music. To extract emotion-related features, it is important to preprocess high-dimensional EEG signals, which may include a significant number of irrelevant features. In literature, the most popular EEG features for emotion detection are fractal dimension (FD), power spectral density (PSD), Higuchi fractal dimension (HFD), multiscale fractal dimension (MFD), spectrograms, and the wavelet transform (WT) [18]. Some of the methods and tools used to preprocess EEG and obtain these features are [18] the Higuchi algorithm [33] for FD, the STFT, WT, and the Hilbert-Huang Transform (HHT) for PSD, and PyEEG [5] for HFD. Wavelet domain features can be extracted using the STFT, high-order crossing analysis (HOC), and hybrid adaptive filtering (HAF). To remove eye movements, facial muscle movements, heartbeats, and other distortions, independent component analysis (ICA) and blind source separation (BSS) can be used [18].

Similar to EEG, ECG signals have also been used for MER. Hsu et al. [35] developed the sequential forward floating selection-kernel-based class separability (SFFS-KBCS) feature selection algorithm and also utilized generalized discriminant analysis (GDA) in order to select significant ECG features for emotion detection. Furthermore, Naji et al. [70] calculated several features using the RR (time intervals between consecutive R peaks) time frames extracted from the

---

[3]https://opencv.org/ (Accessed 15 November 2024).
[4]https://pypi.org/project/dlib/ (Accessed 15 November 2024).

Category	Features
Statistical Features	Mean, Median, Standard Deviation, Range
Time Domain Features	Mean Absolute Value of 1st/2nd Differences of Raw/Normalized Signals Fractal Dimension (FD) of EEG
Frequency Domain Features	Low and High Frequency (LF, HF), Ratio LF/HF
Signal-Specific Features	Fractal Dimension (FD), Power Spectral Density (PSD) Higuchi Fractal Dimension (HFD), Multiscale Fractal Dimension (MFD) Galvanic Skin Response (GSR), Heart Rate Variability (HRV)

Table 5. Categorization of physiological features.

ECG signal. Some of these features are statistical features (e.g. mean and standard deviation), nonlinear features (e.g. sample entropy), and triangular phase space mapping.

Hu et al. [37] defined the relationship between music-induced emotions and physiological signals. They collected electrodermal activity (EDA), blood volume pulse (BVP), inter-beat interval (IBI), heart rate (HR), and skin temperature (TEMP) data and obtained a set of features that can be used in MER (see Table 5 for a summary of these and other physiological features).

3.1.6 Textual Features. Beyond lyrics, other forms of textual data, such as user reviews, social media comments, and textual descriptions of music, have been employed to gauge emotional responses to music. These texts often capture subjective experiences and sentiments of listeners, which can be analyzed using NLP techniques to infer the emotional impact of music. These features are also extracted similarly to other textual data discussed above [106].

3.1.7 Metadata and Contextual Features. Methods employing metadata and contextual information such as song title and artist name as a modality for their final prediction, typically use pretrained models such as BERT to extract the features [117].

3.2 Feature Processing

After features have been extracted (Stage 2) from the raw input data (Stage 1), they typically need to be further processed (Stage 3) before they can be used effectively for final emotion prediction (Stage 4). There are three main feature processing approaches that researchers have used up to now (Fig. 3), which we summarize in this section.

3.2.1 Approach 1: Feature Concatenation. The most straightforward approach is to concatenate the modality-specific features into a unified feature set. Previous research has explored various techniques to accomplish this. Some studies employ a simple vector representation to merge all features within the same space [60], while others adopt more advanced methods, such as utilizing a bimodal deep Boltzmann machine (DBM) for feature fusion [39]. This approach is related to what is generally called early fusion in the deep learning literature [43, 78].

3.2.2 Approach 2: Modality-Specific Feature Processing. An alternative approach is to train modality-specific models and then combine their outputs. There are two variants of this approach, which we refer to as Approach 2-A and Approach 2-B (Fig. 3). In Approach 2-A, each modality-specific model produces an embedding, which is used as input to the final multimodal emotion prediction model. This is known as intermediate fusion [43, 78]. Most works using

this approach train unimodal models, taking the extracted features as input, and then the last few layers of the models are removed to obtain the embeddings [20]. Alternatively, the unimodal models can be used directly as in Approach 2-B, providing unimodal predictions that can be aggregated to produce a final multimodal prediction. This situation corresponds to late fusion [43, 78].

3.2.3 Approach 3: Cross-Modal Feature Processing. The most sophisticated approach is to perform cross-modal feature processing. This concept is increasingly adopted by the field of MER and is especially gaining popularity in MMER. Generally speaking, cross-modal processing combines inputs from different modalities to create a single output. This output is then used as the input for the final classifier, regressor, or fusion model. Unlike multimodal processing, which focuses on integrating multiple types of data to improve overall model performance, cross-modal processing emphasizes how interactions between different modalities can enhance or influence the processing of each data type, creating a more cohesive and contextually aware representation. This takes the concept of intermediate fusion to a higher level.

Recent works [97, 105, 117] have used cross-modal processing to assess the interaction between audio and lyrics. The input consists of pairs of lyric and audio fragments, where each input is typically a single sentence of lyrics and the corresponding audio, and the output is an emotion feature vector. A key element of cross-modal processing in these works is the use of an emotion long-short-term memory (E-LSTM) cell, an enhancement of the traditional LSTM, which processes the next lyric-audio pair along with the emotion vector of the past pair to maintain a consistent emotional state with respect to the previous interaction and to avoid the emotion independence between the two channels. The historical emotion vector is designed to preserve intense emotional information, while the current emotion vector updates relatively weaker emotions. This combined emotion vector is iteratively updated with the historical emotion vector, which helps decide which emotional levels in the song should be maintained and which past emotions need to be revised (see [97] for detailed mathematical formulations of the operations in this process).

3.3 Emotion Prediction

Given the processed features (Stage 3), the final task (Stage 4) is to predict the emotion they convey. The methods used to accomplish this task depend on the form in which the processed features are presented. As alluded to in the previous section, this corresponds to different fusion strategies (Fig. 5), which we elaborate on here as we discuss their usage in ultimate emotion prediction (see Table 6 for an overview of state-of-the-art MMER methods and performance).

3.3.1 Emotion Prediction by Feature-Level Fusion. Feature-level fusion, also known as early fusion, involves combining all extracted features from different modalities into a single high-dimensional feature vector (Approach 1). This unified feature vector is then used to train a single classifier such as support vector machines (SVMs) [60, 104], auto encoders [119], or CNNs [81, 118] for emotion prediction. A common approach to feature-level fusion is feature concatenation, where features from modalities such as audio and lyrics are merged into one vector and fed into a classification model. As this creates a multimodal feature space, challenges arise in integrating heterogeneous features from different modalities into a cohesive representation [108]. Researchers have explored various techniques to address these challenges, including dimensionality reduction and normalization, which aim to better align the feature vectors of varying modalities. Dimensionality reduction methods such as principal component analysis (PCA) aim to reduce the number of features while retaining the most important variance in the data, and normalization methods such as z-score normalization and min-max scaling are often applied to ensure that features from different modalities are on a similar scale, which improves the performance of classification algorithms [55]. Despite these efforts, direct concatenation can
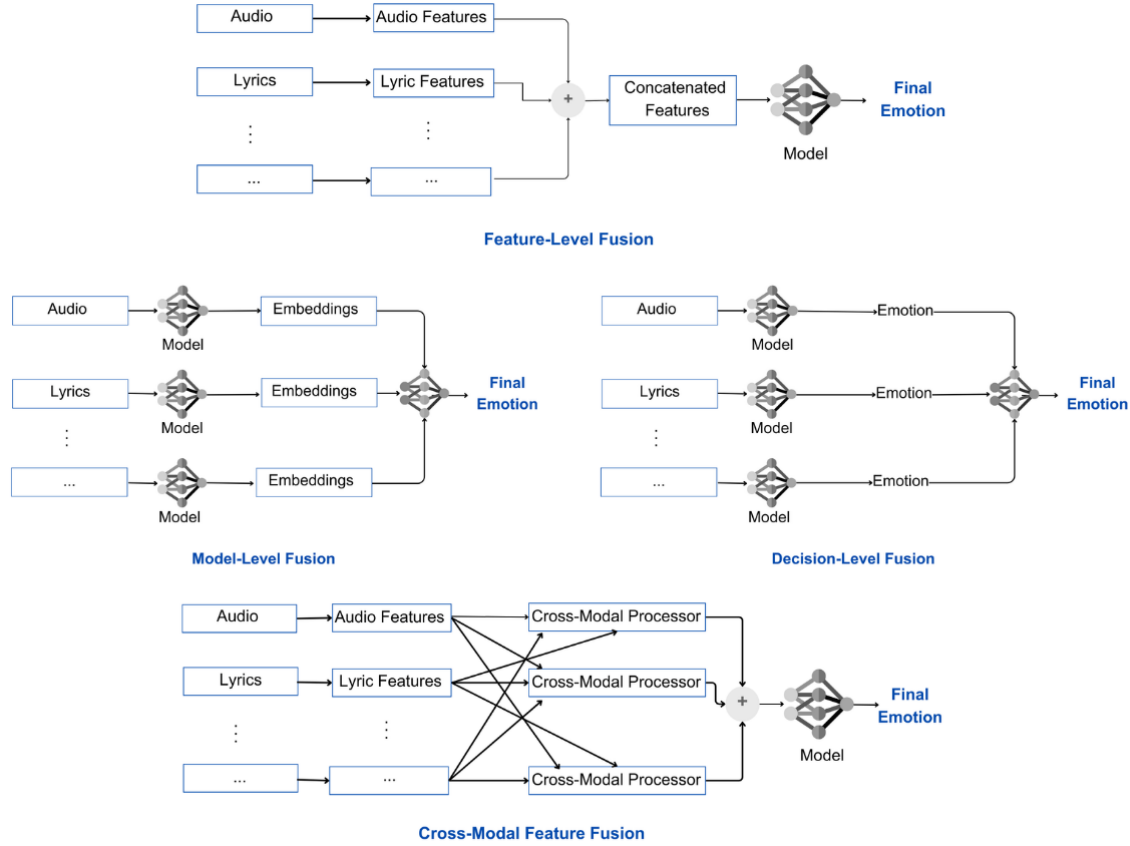
Manuscript

Fig. 5. Comparison of fusion methods in music emotion prediction.

sometimes fail to preserve critical emotional information due to the inherent differences between the representations of modal features [13].

3.3.2 Emotion Prediction by Decision-Level Fusion. Decision-level fusion, also known as late fusion, combines the outputs of unimodal classifiers or regressors (Approach 2-B). These outputs can be aggregated using algebraic combination rules such as "Min," "Max," or "Sum" [13, 40, 63, 73, 77, 112]. Methods in this category can be further divided into linear probability fusion (LPF) and stacking ensemble learning (SEL) [13]. While LPF is widely used in general machine learning applications [56], SEL is more prevalent in the MMER literature [13, 77]. Advanced techniques for decision-level fusion include late fusion subtask merging (LFSM), which integrates audio and lyrics for multimodal emotion categorization [40], and stacking, which assigns distinct weights to each category or modality for better integration [63, 73]. Recent advancements have focused on improving decision-level fusion by employing neural networks for weighting outputs. However, this approach has shown limited efficacy [43]. Instead, stacking methods that combine the outputs of multiple models without fully fusing them have been shown to enhance performance [112].

To obtain unimodal emotion predictions, whether as final output or as input to multimodal models, various methods have been used. For audio, earlier studies have used SVMs, random forests, and logistic regression [39, 60, 104, 108],

while nowadays most audio-based MER is done using deep learning methods such as CNNs and RNNs. Others have experimented with combining a CNN and an RNN in a single model [45]. Continuous audio-based MER is often done by a combination of CNNs and LSTMs [14, 81, 112]. Furthermore, bidirectional LSTMs (BiLSTMs) with attention layers have been used [112]. For audio classification, CNNs have been found to outperform BiGRU models [80]. Some researchers have leveraged transfer learning by using pretrained models on large-scale audio datasets, which are then fine-tuned for specific emotion recognition tasks [13, 31]. Building on this, fusion models have combined CNNs for SF extraction and LSTMs for TF extraction to enhance feature representation [13, 112], while convolutional autoencoders with frequency and time masking techniques have been used to emphasize important features and reduce noise [32]. Researchers have also adopted self-attention mechanisms in RNNs which dynamically weigh the importance of different audio segments [77], and end-to-end learning approaches like SampleCNN to process raw audio signals directly using very small filters [32]. Additionally, they have incorporated multiview neural networks to integrate multiple perspectives of the audio signal [112].

Although audio has thus far been the most prominent modality in MER, other modalities are gaining popularity. Earlier studies on lyric-based MER largely utilized traditional classifiers such as SVMs [60, 104, 108]. However, these methods often fall short in capturing the deeper contextual and semantic nuances of lyrics, which are crucial for accurate emotion recognition [14]. More recent research in lyric-based MER has shifted to deep learning models (Table 6). CNNs, RNNs, and LSTMs are utilized for their ability to handle sequential data and maintain context over long text spans. These models identify and retain emotional indicators spread across multiple lines of lyrics. Neural networks such as the restricted Boltzmann machine (RBM) are used for unsupervised feature learning for lyric classification [119]. Additionally, the BERT model, with its transformer-based architecture, is used to understand the bidirectional context of words, enhancing the accuracy of emotion detection in lyrics [14, 117]. To our knowledge, there exist no works using Approach 2-B with video. MIDI has also been used for training MER models using symbolic features derived from it. Notably, some studies have employed an SVM as the MER model [90], while others have utilized an RNN, specifically a BiGRU [118]. Physiological signals such as EEG have been used with an SVM model [90] and electrodermal activity with a CNN for MER [109]. We are not aware of studies that have used textual data and metadata/contextual data for unimodal emotion recognition.

3.3.3   Emotion Prediction by Model-Level Fusion. Model-level fusion, also referred to as middle-level or intermediate fusion, involves utilizing embeddings (Approach 2-A) or predictions derived from modality-specific models (Approach 2-B) to train a final emotion prediction model. This bridges feature-level and decision-level fusion, offering a balance between combining raw features and final predictions. Techniques for model-level fusion include combining embeddings generated by modality-specific models, such as those for audio and lyrics, into a shared representation for emotion recognition. Additionally, methods like the Hough voting mechanism within the Hough forest model have been explored, effectively integrating cues from multiple modalities to predict music emotions [104]. This allows modalities to contribute jointly while maintaining their distinct representations, thereby enhancing emotion prediction performance [20].

3.3.4   Emotion Prediction by Cross-Modal Fusion. Cross-modal feature fusion has emerged as a promising approach in MMER, gaining traction in more recent studies [97, 105, 117]. This approach mainly aims at modeling and exploiting interaction between modalities at the feature level, effectively allowing the model to incorporate complementary information from different modalities. Unlike traditional fusion strategies, cross-modal fusion insists on direct processing of shared and unique aspects of each modality, fostering richer and more holistic representation.

Manuscript

Cross-modal processing typically involves aligning and integrating features from different modalities in a shared latent space, where relationships between modalities can be explored. Techniques such as attention mechanisms [117] and graph-based methods are often used to model dependencies and interactions between modalities. For example, attention mechanisms dynamically weigh the importance of specific features in one modality based on their relevance to another, enhancing the model's ability to prioritize meaningful interactions. Similarly, graph-based approaches construct cross-modal graphs to represent and learn complex relationships between modalities, facilitating a structured fusion process. Although graph-based approaches are not commonly applied in MMER, they have been widely used in other areas of multimodal research such as video understanding [116], and audio-visual event localization [103].

Reference	Year	Modalities	Datasets	Methods	Approach	Performance
[60]	2008	Audio, Lyrics	Own (last.fm + LyricWiki)	SVM	1	Accuracy: 92.40%
[108]	2008	Audio, Lyrics	Own	SVM	2-A	Accuracy: 73.32% Valence 78.03% Arousal
[72]	2013	Audio, Lyrics, MIDI	Own	SVM	1	Accuracy: 61.20%
[104]	2015	Audio, Lyrics	Own	SVM	1	Accuracy: 59.9%
[39]	2016	Audio, Lyrics	Million Song DB	SVM	2-A	Accuracy: 72.18%
[45]	2017	Audio, Lyrics	Own	CNN+RNN, CNN	2-A	Accuracy: 80.46%
[90]	2017	MIDI, EEG	Own	SVM	2-B	Accuracy: 87.21%
[20]	2018	Audio, Lyrics	Own (Million Song DB + Deezer Catalog)	CNN , LSTM	2-A	R^2: 0.219 Valence 0.232 Arousal
[119]	2019	Audio, Lyrics	Own	Bimodal Deep Auto Encoder	1	Accuracy: 79.9% Pleasure 81.5% Arousal 67.7% Dominance
[63]	2020	Audio, Lyrics	Own	LSTM , BERT	2-B	Accuracy: 79.62%
[13]	2021	Audio, Lyrics	Million Songs DB	Multifeature Combined Classifier (CNN-LSTM Based)	2-A	Accuracy: 78.2%
[73]	2022	Audio, Video, Facial Expression	Own	CNN + OpenCV	2-A	Accuracy: 74%
[77]	2022	Audio, Lyrics	MoodyLyrics	CNN , BERT	2-B	Accuracy: 94.58%
[112]	2022	Audio, Lyrics	Own	CNN-LSTM	2-A	Accuracy: 78.2%
[14]	2022	Audio, Videos	Own	CNN-LSTM	2-A	Accuracy: 77.9%
[117]	2022	Audio, Lyrics, Track Name, Artist	Million Songs DB + MoodyLyrics	CNN , BERT	3	R^2: 0.306 Valence 0.311 Arousal
[95]	2022	Audio, Lyrics	Own	CNN + NN	2-B	Accuracy: 90.89%
[109]	2022	Audio, Electrodermal Activity Signals	PMEmo + DEAP+ AMIGOS	RTCAN-1D	1	Accuracy: 79.68% Valence 83.76% Arousal

Reference	Year	Modalities	Datasets	Methods	Approach	Performance
[80]	2023	Audio, Lyrics	MoodyLyrics	BiGRU + Attention	2-A	Accuracy: 77.94%
[118]	2023	Audio, MIDI	EMOPIA	BiGRU , CNN	1	Accuracy: 69.2%
[81]	2023	Audio, Lyrics	Own (Indonesian)	CNN-LSTM , XLNet	1	Accuracy: 80.56%
[97]	2024	Audio, Lyrics	DEAM + FMA	CNN , ALBERT	3	Accuracy: 49.68% DEAM 49.54% FMA
[105]	2024	Audio, Lyrics, Song Structure	Own	OpenSMILE + BERT	3	RMSE: 0.162 Valence 0.147 Arousal

Table 6. Summary of literature on MMER with performance.

4 Current Status and Future Directions

The field of MMER has seen significant advancements in recent years. Nevertheless, there is still much room for further performance improvement (Table 6). To assist ongoing developments, we summarize the current trends (Section 4.1) and challenges (Section 4.2), as well as potential future research directions (Section 4.3).

4.1 Current Trends

Analyzing the recent literature, we observed some noticeable development trends in MMER, especially the growing availability of multimodal datasets and the shift from traditional machine learning methods to increasingly advanced multimodal deep learning methods.

4.1.1 Multimodal Datasets. With MMER gaining popularity, existing MER datasets have started to include data of different modalities, and new datasets are increasingly multimodal from the outset (Table 3), including video, physiological signals, and textual data such as comments and metadata. Moreover, due to the dynamic nature of emotion in music, the field is now moving away from static processing, as it may not yield optimal and sufficiently detailed predictions for a given application. Therefore, datasets such as MERP [58], DEAM [3], PMEmo [111], and MuVi[17] include dynamic annotation rather than static annotation. Also, the advancement of techniques such as RNN models has made dynamic processing more efficient, further accelerating the emergence of dynamically annotated datasets. One of the key driving factors in recent efforts to expand and create multimodal datasets is that they allow for the development of better emotion prediction methods. For example, in a study by Liu and Tan [63], the highest accuracies achieved by unimodal methods were 70.6% for audio and 62.9% for lyrics, while multimodal methods reached 79.2%. Another study, by Chen and Li [13], reported 68.1% and 74.2% accuracy for audio and lyric classification, respectively, compared to 78.2% accuracy for the multimodal model.

4.1.2 Multimodal Methods. Over the years, the majority of MMER papers have considered audio and lyrics as the primary modalities (Table 6), and the developed methods have progressed from traditional machine learning techniques such as SVMs to more and more sophisticated deep learning models based on CNNs, RNNs (in particular LSTMs), and recently also Transformers [13, 14, 80, 81, 112]. In addition, various fusion strategies have been developed to effectively combine the two modalities. The highest accuracy of 94.58% for classification to date has been achieved by employing a CNN for audio analysis, BERT for lyrics analysis, and using late fusion of the two [77]. Other modalities such as

metadata, song structure, MIDI, and video have also been incorporated into MMER systems [72, 105, 117]. A common approach is to combine audio with video, although the highest accuracy reported thus far is only 77.9% [14]. Modalities such as MIDI and physiological signals have as yet seen limited usage, with SVMs and CNNs being the most popular methods for prediction. The use of metadata has also received little attention to date. Processing of such data is typically done using models such as BERT and ALBERT [117].

Beyond traditional machine learning and deep learning models, some researchers have explored alternative technologies like the OpenSMILE[5] toolkit, which can process a variety of modalities including audio, visual data, and physiological signals, offering a versatile approach to MMER. Another very important development highlighted in this survey is the use of cross-modal processing, which allows for more comprehensive and accurate emotion detection by enabling methods to process and correlate information from different sensory inputs simultaneously. Leveraging diverse modalities, cross-modal processing enhances the ability of MMER methods to understand complex emotional cues, leading to improved performance in various applications, from virtual assistants to mental health monitoring tools [55].

4.2 Current Challenges

Despite recent advancements and the growing popularity of MMER, it is still quite far from achieving human-like performance (Table 6). One of the main challenges is the subjectivity of music experience. Different people can feel different emotions in a single song. Even the same person can feel different emotions for the same song depending on their mood [52]. According to research, emotion can be influenced by a variety of factors, including culture, genre preference, age, level of music expertise, gender, and even the weather. Therefore, some studies have considered the profile data of annotators and listeners when making predictions [58], and more recent works [29] even considered the time and weather during listening to the song when making predictions. Accurately classifying music using categorical emotion models is very difficult. In dimensional emotion models, particularly the widely used valence-arousal model, each quadrant has an extensive range of emotions (Fig. 1), making precise annotation ambiguous and challenging. It might be beneficial to introduce a new emotion model specifically for the purpose of MER.

Another primary issue is the limited availability of datasets. The majority of datasets, such as DEAP, emoMV, 4Q [57, 71, 91] do not include complete song audios due to copyright restrictions. Furthermore, the number of songs or song excerpts is very limited, ranging from 100s to 1,000s. Some more recent datasets including Audioset [27], MTG-Jamendo [9], MuSe [1], and Music4all [74] have 17k, 55k, 90k, and 109k songs, respectively (Table 3), but they contain only a single modality (mostly audio only), making them unsuitable for MMER. Another limitation of current datasets is that they tend to focus on a single genre of music, resulting in a lack of diversity in the songs included. For instance, DEAM mainly includes rock and techno music, and MSD (Million Song Dataset) [7] consists primarily of pop music. Thus, there is an urgent need for larger and more diverse multimodal datasets for MMER. This would also open the door for the creation of state-of-the-art benchmarks for MMER, which are currently lacking.

A continuing problem in MER in general is the insufficient use of music theory and concepts, such as considering the mode the music is written in and loudness variations, in the process of recognizing emotions. It is known that concepts like tempo, pitch, and rhythm can convey emotion (Section 2.1), but thus far only a small number of studies have made use of these.

---

[5]https://www.audeering.com/research/opensmile/ (Accessed 15 November 2024).

4.3 Future Directions

To deal with dataset limitations, one option is to employ unsupervised learning methods such as autoencoders [119] as an alternative to the supervised approaches predominantly employed in prior studies. Unsupervised learning enables the utilization of custom unlabeled datasets, which researchers can generate using various APIs and tools. For example, the SoundCloud API[6] can facilitate audio data collection, Lyrics.ovh[7] and MusixMatch[8] can provide lyrics data, and video data can be sourced using Python libraries such as PyTube[9]. This approach allows researchers to overcome reliance on preexisting datasets, offering greater flexibility and customization in data collection to meet specific research objectives. Additionally, incorporating annotators from different cultures when creating datasets can provide a diverse perspective on music emotions. Another approach to overcome the dataset limitation problem is to use transfer learning. Although transfer learning is widely utilized and has produced excellent results in other domains such as computer vision [44], it has seen very little use in the field of MER, and should be explored in the future.

Another potential future direction for MER is to incorporate different modalities other than only audio and lyrics [107]. Researchers can acquire a better understanding of the listener's emotional response to music by combining physiological data like EEG signals, heart rate variability, and skin conductance. These physiological indicators offer a more objective assessment of emotional states, complementing personal judgements based on audio and lyrical content. MIDI is also an invaluable modality to incorporate. To comprehend the structural and expressive elements of a piece, one can analyze the precise symbolic information that MIDI provides about the music, such as note pitches, intervals, and dynamics. Furthermore, as noted in the previous section, involving music theory and concepts is a promising future direction. Integrating these many senses can result in a more complete and comprehensive understanding of emotions in music [55].

Real-time MER is another promising future direction with many possible applications. Currently, many MER models are neither user-friendly nor optimized for quick execution, demanding GPUs and considerable processing time. However, the advancement of real-time technology has the potential to transform multiple sectors. Real-time emotion detection in therapeutic settings can improve emotional regulation techniques and mental health treatments by giving instant feedback to therapists and patients. Real-time analysis could help mood guide playlist systems improve user experience by dynamically changing song choices to better fit the listener's current emotional state. Furthermore, real-time emotion identification could be used in commercial applications to deliver more effective and personalized marketing content, such as targeted advertising [89]. Real-time emotion recognition in music has great potential for both private and public applications as technology develops and these models become more affordable and effective.

Creating real-time MMER applications is quite challenging, primarily due to high computational demands. Combining and synchronizing data streams like audio, video, and text requires advanced algorithms and significant processing power [4]. This complexity often hampers real-time processing. Additionally, the substantial resource demands such as increased memory, processing power, and energy consumption pose further difficulties. This is particularly problematic for portable or embedded systems, where resources are limited. As a result, implementing real-time MMER on less powerful devices can be impractical, hindering broader applications [87]. However, these can be mitigated by adopting efficient data fusion methods. Developing lightweight algorithms that effectively combine multiple modalities with minimal computational overhead is a key strategy. Additionally, leveraging parallel processing techniques or utilizing

---

[6]https://developers.soundcloud.com/docs/api/guide (Accessed 15 November 2024).
[7]https://lyricsovh.docs.apiary.io (Accessed 15 November 2024).
[8]https://developer.musixmatch.com (Accessed 15 November 2024).
[9]https://pytube.io/ (Accessed 15 November 2024).

specialized hardware, such as GPUs, can significantly enhance the system's ability to handle multiple data streams simultaneously without compromising speed. These approaches help maintain the real-time functionality of the system while addressing the inherent computational and resource demands.

5 Conclusion

MMER is an upcoming and developing field in the area of music information retrieval. Although unimodal MER has achieved exceptional performances, MMER has not reached human-like accuracy yet. From the studies surveyed in this paper, we conclude that the majority of MMER methods are limited to using audio and lyrics as the primary modalities. Incorporating other modalities such as video, MIDI, physiological signals, metadata and other (con)textural data may provide additional information and improve emotion recognition performance. Moreover, as most works employ machine learning and deep learning methods to train emotion prediction models, primary factors limiting progress in MMER are the absence of a precise emotional model, limited datasets, and a lack of state-of-the-art benchmarks. Thus, future research should focus on the integration of a wider variety of modalities, the exploration of new deep learning architectures, and the development of comprehensive evaluation frameworks. Notwithstanding current challenges and limitations, MMER holds significant promise. As the field continues to grow, MMER will not only bridge the gap between human-like emotional understanding and machine perception, but also introduce new opportunities for enhancing user experiences and emotional intelligence in technology, revolutionizing human-computer interaction, multimedia content production, medical treatment, recommendation systems, and commercial applications.

References

[1] Christopher Akiki and Manuel Burghardt. 2021. MuSe: The musical sentiment dataset. *Journal of Open Humanities Data* 7 (2021), 10. https://doi.org/10.5334/johd.33

[2] Anna Aljanaki, Frans Wiering, and Remco C. Veltkamp. 2016. Studying emotion induced by music through a crowdsourcing game. *Information Processing & Management* 52, 1 (2016), 115–128. https://doi.org/10.1016/j.ipm.2015.03.004

[3] Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. 2017. Developing a benchmark for emotional analysis of music. *PLOS ONE* 12, 3 (2017), e0173392. https://doi.org/10.1371/journal.pone.0173392

[4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2017. Multimodal Machine Learning: A Survey and Taxonomy. arXiv:1705.09406 [cs.LG] https://arxiv.org/abs/1705.09406

[5] Forrest Sheng Bao, Xin Liu, and Christina Zhang. 2011. PyEEG: An open source Python module for EEG/MEG feature extraction. *Computational Intelligence and Neuroscience* 2011, 1 (2011), 406391. https://doi.org/10.1155/2011/406391

[6] Lisa Feldman Barrett. 2006. Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review* 10, 1 (2006), 20–46. https://doi.org/10.1207/s15327957pspr1001_2 PMID: 16430327.

[7] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. 2011. The million song dataset. In *International Society for Music Information Retrieval Conference (ISMIR)*. International Society for Music Information Retrieval, Miami, Florida, USA, 591–596. https://doi.org/10.7916/D8NZ8J07

[8] Dmitry Bogdanov, Xabier Lizarraga-Seijas, Pablo Alonso-Jiménez, and Xavier Serra. 2022. MusAV: A dataset of relative arousal-valence annotations for validation of audio models. In *International Society for Music Information Retrieval Conference (ISMIR)*. International Society for Music Information Retrieval, Bengaluru, India, 650–658. https://doi.org/10.5281/zenodo.7316746

[9] Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. 2019. The MTG-Jamendo dataset for automatic music tagging. In *International Conference on Machine Learning*. Semantic Scholar, Long Beach, California, USA, 1–3. https://api.semanticscholar.org/CorpusID:196187495

[10] Cibele Maia Burke. 2017. A comparative study of perspectives in musical structural features and emotional stimuli. In *Honors Theses*. Eastern Kentucky University, Kentucky, USA, 1–21. https://api.semanticscholar.org/CorpusID:67822675

[11] Nathan R. Carr, Kirk N. Olsen, and William Forde Thompson. 2023. The perceptual and emotional consequences of articulation in music. *Music Perception* 40, 3 (2023), 202–219. https://doi.org/10.1525/mp.2023.40.3.202

[12] Vybhav Chaturvedi, Arman Beer Kaur, Vedansh Varshney, Anupam Garg, Gurpal Singh Chhabra, and Munish Kumar. 2021. Music mood and human emotion recognition based on physiological signals: A systematic review. *Multimedia Systems* 28 (2021), 21–44. https://api.semanticscholar.org/CorpusID:234871442

[13] Changfeng Chen and Qiang Li. 2020. A multimodal music emotion classification method based on multifeature combined network classifier. *Mathematical Problems in Engineering* 2020, 1 (2020), 4606027. https://doi.org/10.1155/2020/4606027

[14] Wenwen Chen. 2022. A novel long short-term memory network model for multimodal music emotion analysis in affective computing. *Journal of Applied Science and Engineering* 26, 3 (2022), 367–376. https://doi.org/10.6180/jase.202303_26(3).0008

[15] Yu-An Chen, Yi-Hsuan Yang, Ju-Chiang Wang, and Homer Chen. 2015. The AMG1608 dataset for music emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, South Brisbane, Queensland, Australia, 693–697. https://doi.org/10.1109/ICASSP.2015.7178058

[16] Keunwoo Choi, György Fazekas, Mark B. Sandler, and Kyunghyun Cho. 2017. Convolutional recurrent neural networks for music classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, New Orleans, Louisiana, USA, 2392–2396. https://doi.org/10.1109/ICASSP.2017.7952585

[17] Phoebe Chua, Dimos Makris, Dorien Herremans, Gemma Roig, and Kat Agres. 2022. Predicting emotion from music videos: Exploring the relative contribution of visual and auditory information to affective responses. arXiv:2202.10453 https://doi.org/10.48550/arXiv.2202.10453

[18] Xu Cui, Yongrong Wu, Jipeng Wu, Zhiyu You, Jianbing Xiahou, and Menglin Ouyang. 2022. A review: Music-emotion recognition and analysis based on EEG signals. *Frontiers in Neuroinformatics* 16 (2022), 1–17. https://doi.org/10.3389/fninf.2022.997282

[19] K. Dakshina and Rajeswari Sridhar. 2014. LDA based emotion recognition from lyrics. In *Advanced Computing, Networking and Informatics*, Vol. 1. Springer International Publishing, Cham, Switzerland, 187–194. https://doi.org/10.1007/978-3-319-07353-8_22

[20] Rémi Delbouys, Romain Hennequin, Francesco Piccoli, Jimena Royo-Letelier, and Manuel Moussallam. 2018. Music mood detection based on audio and lyrics with deep neural net. arXiv:1809.07276 https://doi.org/10.48550/arXiv.1809.07276

[21] Wan Ding, Mingyu Xu, Dongyan Huang, Weisi Lin, Minghui Dong, Xinguo Yu, and Haizhou Li. 2016. Audio and face video emotion recognition in the wild using deep neural networks and small datasets. In *ACM International Conference on Multimodal Interaction (ICMI)*. Association for Computing Machinery, New York, NY, USA, 506–513. https://doi.org/10.1145/2993148.2997637

[22] Isabela Dogaru, Adrian Furnham, and Alastair McClelland. 2024. Understanding how the presence of music in advertisements influences consumer behaviour. *Acta Psychologica* 248 (2024), 104333. https://doi.org/10.1016/j.actpsy.2024.104333

[23] Darren Edmonds and João Sedoc. 2021. Multi-emotion classification for song lyrics. In *Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, Kerrville, Texas, USA, 221–235. https://aclanthology.org/2021.wassa-1.24

[24] Tuomas Eerola and Jonna K. Vuoskoski. 2011. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music* 39, 1 (2011), 18–49. https://doi.org/10.1177/0305735610362821

[25] Paul Ekman. 1971. Universals and cultural differences in facial expressions of emotion. In *Nebraska Symposium on Motivation*. University of Nebraska Press, Nebraska, USA, 207–283. https://psycnet.apa.org/record/1973-11154-001

[26] Zhouyu Fu, Guojun Lu, Kai Ting, and Dengsheng Zhang. 2011. A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia* 13, 2 (2011), 303–319. https://doi.org/10.1109/TMM.2010.2098858

[27] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, New Orleans, Louisiana, USA, 776–780. https://doi.org/10.1109/ICASSP.2017.7952261

[28] Juan Sebastián Gómez-Cañón, Nicolás Gutiérrez-Páez, Lorenzo Porcaro, Alastair Porter, Estefanía Cano, Perfecto Herrera-Boyer, Aggelos Gkiokas, Patricia Santos, Davinia Hernández-Leo, Casper Karreman, and Emilia Gómez. 2022. TROMPA-MER: An open dataset for personalized music emotion recognition. *Journal of Intelligent Information Systems* 60, 2 (2022), 549–570. https://doi.org/10.1007/s10844-022-00746-0

[29] Vadim Grigorev, Jiayu Li, Weizhi Ma, Zhiyu He, Min Zhang, Yiqun Liu, Ming Yan, and Ji Zhang. 2024. SiTunes: A situational music recommendation dataset with physiological and psychological signals. In *Conference on Human Information Interaction and Retrieval (CHIIR)*. Association for Computing Machinery, New York, NY, USA, 417–421. https://doi.org/10.1145/3627508.3638343

[30] Donghong Han, Yanru Kong, Han Jiayi, and Guoren Wang. 2022. A survey of music emotion recognition. *Frontiers of Computer Science* 16 (2022), 166335. https://doi.org/10.1007/s11704-021-0569-4

[31] Xiao Han, Fuyang Chen, and Junrong Ban. 2023. Music emotion recognition based on a neural network with an inception-gru residual structure. *Electronics* 12, 4 (2023), 978. https://doi.org/10.3390/electronics12040978

[32] Na He and Sam Ferguson. 2022. Music emotion recognition based on segment-level two-stage learning. *International Journal of Multimedia Information Retrieval* 11, 3 (2022), 383–394. https://doi.org/10.1007/s13735-022-00230-z

[33] T. Higuchi. 1988. Approach to an irregular time series on the basis of the fractal theory. *Physica D: Nonlinear Phenomena* 31, 2 (1988), 277–283. https://doi.org/10.1016/0167-2789(88)90081-4

[34] Mohammad Hossin and Sulaiman M.N. 2015. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process* 5, 2 (2015), 1–11. https://doi.org/10.5121/ijdkp.2015.5201

[35] Yu-Liang Hsu, Jeen-Shing Wang, Wei-Chun Chiang, and Chien-Han Hung. 2020. Automatic ECG-based emotion recognition in music listening. *IEEE Transactions on Affective Computing* 11, 1 (2020), 85–99. https://doi.org/10.1109/TAFFC.2017.2781732

[36] Xiao Hu, Fanjie Li, and Ruilun Liu. 2022. Detecting music-induced emotion based on acoustic analysis and physiological sensing: A multimodal approach. *Applied Sciences* 12, 18 (2022), 9354. https://doi.org/10.3390/app12189354

[37] Xiao Hu, Fanjie Li, and Jeremy Ng. 2018. On the relationships between music-induced emotion and physiological signals. In *International Society for Music Information Retrieval Conference (ISMIR)*. International Society for Music Information Retrieval, Paris, France, 362–369. https://archives.ismir.net/ismir2018/paper/000115.pdf

[38] Jingyue Huang, Ke Chen, and Yi-Hsuan Yang. 2024. Emotion-driven Piano Music Generation via Two-stage Disentanglement and Functional Representation. arXiv:2407.20955 [cs.SD] https://arxiv.org/abs/2407.20955

[39] Moyuan Huang, Wenge Rong, Tom Arjannikov, Nan Jiang, and Zhang Xiong. 2016. Bi-modal deep Boltzmann machine based musical emotion classification. In *Artificial Neural Networks and Machine Learning (ICANN)*. Springer International Publishing, Cham, Switzerland, 199–207. https://doi.org/10.1007/978-3-319-44781-0_24

[40] Álvaro Huertas-García, Helena Liz, Guillermo Villar-Rodríguez, Alejandro Martín, Javier Huertas-Tato, and David Camacho. 2022. AIDA-UPM at SemEval-2022 Task 5: Exploring multimodal late information fusion for multimedia automatic misogyny identification. In *International Workshop on Semantic Evaluation (SemEval)*. Association for Computational Linguistics, Seattle, USA, 771–779. https://doi.org/10.18653/v1/2022.semeval-1.107

[41] Hsiao-Tzu Hung, Joann Ching, Seungheon Doh, Nabin Kim, Juhan Nam, and Yi-Hsuan Yang. 2021. EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation. arXiv:2108.01374 https://arxiv.org/abs/2108.01374

[42] Arefin Huq, Juan Pablo Bello, and Robert Rowe. 2010. Automated music emotion recognition: A systematic evaluation. *Journal of New Music Research* 39, 3 (2010), 227–244. https://doi.org/10.1080/09298215.2010.513733

[43] Mazhar Hussain, Mattias O'Nils, Jan Lundgren, and Seyed Jalaleddin Mousavirad. 2024. A comprehensive review on deep learning-based data fusion. *IEEE Access* 12 (2024), 180093–180124. https://doi.org/10.1109/ACCESS.2024.3508271

[44] Mohammadreza Iman, Hamid Reza Arabnia, and Khaled Rasheed. 2023. A review of deep transfer learning and recent advancements. *Technologies* 11, 2 (2023), 40. https://doi.org/10.3390/technologies11020040

[45] Byungsoo Jeon, Chanju Kim, Adrian Kim, Dongwon Kim, Jangyeon Park, and Jung-Woo Ha. 2017. Music emotion recognition via end-to-end multimodal neural networks. In *ACM Conference on Recommender Systems (RecSys)*. ACM, Como, Italy, 1–2. https://api.semanticscholar.org/CorpusID:416794

[46] Il-Young Jeong and Kyogu Lee. 2016. Learning temporal features using a deep neural network and its application to music genre classification. In *International Society for Music Information Retrieval Conference (ISMIR)*. International Society for Music Information Retrieval, New York, NY, USA, 434–440. https://archives.ismir.net/ismir2016/paper/000159.pdf

[47] Mukkamala Jitendra and Radhika Yalavarthi. 2020. A review: Music feature extraction from an audio signal. *International Journal of Advanced Trends in Computer Science and Engineering* 9 (2020), 973–980. https://doi.org/10.30534/ijatcse/2020/11922020

[48] Charles Joseph and Sugeeswari Lekamge. 2019. Machine learning approaches for emotion classification of music: A systematic literature review. In *International Conference on Advancements in Computing (ICAC)*. IEEE, Malabe, Sri Lanka, 334–339. https://doi.org/10.1109/ICAC49085.2019.9103378

[49] Patrik N. Juslin. 2001. Communicating emotion in music performance: A review and theoretical framework. In *Music and Emotion: Theory and Research*. Oxford Academic, New York, NY, USA, 309–338. https://doi.org/10.1093/oso/9780192631886.003.0014

[50] Patrik N. Juslin and Petri Laukka. 2004. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research* 33, 3 (2004), 217–238. https://doi.org/10.1080/0929821042000317813

[51] Patrik N. Juslin, Laura S. Sakka, Gonçalo T. Barradas, and Olivier Lartillot. 2022. Emotions, mechanisms, and individual differences in music listening: A stratified random sampling approach. *Music Perception* 40, 1 (2022), 55–86. https://doi.org/10.1525/mp.2022.40.1.55

[52] Patrik N. Juslin and Daniel Västfjäll. 2008. Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences* 31, 5 (2008), 559–575. https://doi.org/10.1017/S0140525X08005293

[53] Chenfei Kang, Peiling Lu, Botao Yu, Xu Tan, Wei Ye, Shikun Zhang, and Jiang Bian. 2023. EmoGen: Eliminating Subjective Bias in Emotional Music Generation. arXiv:2307.01229 [cs.SD] https://arxiv.org/abs/2307.01229

[54] Kathi J Kemper and Suzanne C Danhauer. 2005. Music as therapy. *Southern Medical Journal* 98, 3 (2005), 282–288. https://doi.org/10.1097/01.smj.0000154773.11986.39

[55] Youngmoo Kim, Erik Schmidt, Raymond Migneco, Brandon Morton, Patrick Richardson, Jeffrey Scott, Jacquelin Speck, and Douglas Turnbull. 2010. Music emotion recognition: A state of the art review. In *International Society for Music Information Retrieval Conference (ISMIR)*. International Society for Music Information Retrieval, Utrecht, The Netherlands, 255–266. https://archives.ismir.net/ismir2010/paper/000045.pdf

[56] Josef Kittler, Mr. Hatef, Robert Duin, and Jiri Matas. 1998. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 3 (1998), 226–239. https://doi.org/10.1109/34.667881

[57] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2012. DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing* 3, 1 (2012), 18–31. https://doi.org/10.1109/T-AFFC.2011.15

[58] En Yan Koh, Kin Wai Cheuk, Kwan Yee Heung, Kat R. Agres, and Dorien Herremans. 2023. MERP: A music dataset with emotion ratings and raters' profile information. *Sensors* 23, 1 (2023), 382. https://doi.org/10.3390/s23010382

[59] Olivier Lartillot, Petri Toiviainen, and Tuomas Eerola. 2007. A Matlab toolbox for music information retrieval. In *Annual Conference of the Gesellschaft für Klassifikation*. Springer, Berlin, Heidelberg, 261–268. https://api.semanticscholar.org/CorpusID:17342536

[60] Cyril Laurier, Jens Grivolla, and Perfecto Herrera. 2008. Multimodal music mood classification using audio and lyrics. In *International Conference on Machine Learning and Applications (ICMLA)*. IEEE, San Diego, CA, USA, 688–693. https://doi.org/10.1109/ICMLA.2008.96

[61] Richard S. Lazarus. 1995. Vexing research problems inherent in cognitive-mediational theories of emotion- and some solutions. *Psychological Inquiry* 6, 3 (1995), 183–196. https://doi.org/10.1207/s15327965pli0603_1

[62] Kristen A. Lindquist and Lisa Feldman Barrett. 2008. Constructing emotion: The experience of fear as a conceptual act. *Psychological Science* 19, 9 (2008), 898–903. https://doi.org/10.1111/j.1467-9280.2008.02174.x

[63] Gaojun Liu and Zhiyuan Tan. 2020. Research on multi-modal music emotion classification based on audio and lyrics. In *IEEE Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*. IEEE, Chongqing, China, 2331–2335. https://doi.org/10.1109/ITNEC48623.2020.9084846

[64] Zhiyuan Liu, Wei Xu, Wenping Zhang, and Qiqi Jiang. 2023. An emotion-based personalized music recommendation framework for emotion improvement. *Information Processing & Management* 60, 3 (2023), 103256. https://doi.org/10.1016/j.ipm.2022.103256

[65] Beth Logan. 2000. Mel frequency cepstral coefficients for music modeling. In *International Symposium on Music Information Retrieval (MUSIC IR)*. University of Massachusetts, Plymouth, Massachusetts, USA, 1–2. https://ismir2000.ismir.net/papers/logan_abs.pdf

[66] Pedro Lima Louro, Hugo Redinho, Ricardo Santos, Ricardo Malheiro, Renato Panda, and Rui Pedro Paiva. 2024. MERGE – A bimodal dataset for static music emotion recognition. arXiv:2407.06060 https://arxiv.org/abs/2407.06060

[67] Ricardo Malheiro, Renato Panda, Paulo Gomes, and Rui Pedro Paiva. 2013. Music emotion recognition from lyrics: A comparative study. In *International Workshop on Music and Machine Learning (MML)*. MML, Prague, Czech Republic, 1–4. https://hdl.handle.net/10316/95165

[68] Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* 29, 3 (2013), 436–465. https://doi.org/10.1111/j.1467-8640.2012.00460.x

[69] Pampati Nagaraju and Manchala Sadanandam. 2024. Advancements in motion detection within video streams through the integration of optical flow estimation and 3D-convolutional neural network architectures. *Journal of Electrical Systems* 20, 6 (2024), 2502–2517. https://doi.org/10.52783/jes.3238

[70] Mohsen Naji, Mohammad Firoozabadi, and Parviz Azadfallah. 2013. Classification of music-induced emotions based on information fusion of forehead biosignals and electrocardiogram. *Cognitive Computation* 6 (2013), 241–252. https://doi.org/10.1007/s12559-013-9239-7

[71] Renato Panda, Ricardo Malheiro, and Rui Pedro Paiva. 2018. Musical texture and expressivity features for music emotion recognition. In *International Society for Music Information Retrieval Conference (ISMIR)*. International Society for Music Information Retrieval, Paris, France, 383–391. https://api.semanticscholar.org/CorpusID:53875359

[72] Renato Panda, Ricardo Malheiro, Bruno Rocha, António Oliveira, and Rui Pedro Paiva. 2013. Multi-modal music emotion recognition: A new dataset, methodology and comparative analysis. In *International Symposium on Computer Music Multidisciplinary Research (CMMR)*. Springer Verlag, Marseille, France, 570–582. https://hdl.handle.net/10316/94095

[73] Yagya Raj Pandeya, Bhuwan Bhattarai, and Joonwhoan Lee. 2021. Deep-learning-based multimodal emotion classification for music videos. *Sensors* 21, 14 (2021), 4927. https://doi.org/10.3390/s21144927

[74] Igor André Pegoraro Santana, Fabio Pinhelli, Juliano Donini, Leonardo Catharin, Rafael Biazus Mangolin, Yandre Maldonado e Gomes da Costa, Valéria Delisandra Feltrim, and Marcos Aurélio Domingues. 2020. Music4All: A new music database and its applications. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE, Niteroi, Brazil, 399–404. https://doi.org/10.1109/IWSSIP48289.2020.9145170

[75] Robert Plutchik and Henry Kellerman. 2013. *EMOTION: Theory, Research, and Experience*. Theories of emotion, Vol. 1. Academic press, New York, NY, USA. https://www.sciencedirect.com/book/9780125587013/

[76] Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology* 17, 3 (2005), 715–734. https://doi.org/10.1017/S0954579405050340

[77] Konstantinos Pyrovolakis, Paraskevi Tzouveli, and Giorgos Stamou. 2022. Multi-modal song mood detection with deep learning. *Sensors* 22, 3 (2022), 1065. https://doi.org/10.3390/s22031065

[78] Dhanesh Ramachandram and Graham W. Taylor. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine* 34, 6 (2017), 96–108. https://doi.org/10.1109/MSP.2017.2738401

[79] James Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178. https://doi.org/10.1037/h0077714

[80] Sujeesha A. S. and Rajan Rajeev. 2023. Transformer-based automatic music mood classification using multi-modal framework. *Journal of Computer Science and Technology* 23, 1 (2023), e02. https://doi.org/10.24215/16666038.23.e02

[81] Andrew Sams and Amalia Zahra. 2023. Multimodal music emotion recognition in Indonesian songs based on CNN-LSTM, XLNet transformers. *Bulletin of Electrical Engineering and Informatics* 12, 1 (2023), 355–364. https://doi.org/10.11591/eei.v12i1.4231

[82] Klaus Scherer. 2001. Appraisal considered as a process of multilevel sequential checking. In *Appraisal Processes in Emotion: Theory, Methods, Research*. Oxford University Press, Oxford, UK, 92–120. https://doi.org/10.1093/oso/9780195130072.003.0005

[83] Klaus R. Scherer. 2005. What are emotions? And how can they be measured? *Social Science Information* 44, 4 (2005), 695–729. https://doi.org/10.1177/0539018405058216

[84] Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management* 45, 4 (2009), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

[85] Mohammad Soleymani, Micheal N. Caro, Erik M. Schmidt, Cheng-Ya Sha, and Yi-Hsuan Yang. 2013. 1000 songs for emotional analysis of music. In *ACM International Workshop on Crowdsourcing for Multimedia (CrowdMM)* (Barcelona, Spain) *(CrowdMM '13)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/2506364.2506365

[86] Hannah Strauß, Julia Vigl, Peer-Ole Jacobsen, Martin Bayer, Francesca Talamini, Wolfgang Vigl, Eva Zangerle, and Marcel Zentner. 2024. The Emotion-to-Music Mapping Atlas (EMMA): A systematically organized online database of emotionally evocative music excerpts. *Behavior Research Methods* 56 (2024), 3560–3577. https://doi.org/10.3758/s13428-024-02336-0

[87] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel Emer. 2017. Efficient Processing of Deep Neural Networks: A Tutorial and Survey. arXiv:1703.09039 [cs.CV] https://arxiv.org/abs/1703.09039

[88] Abhishek V. Tatachar. 2021. Comparative assessment of regression models based on model evaluation metrics. *International Journal of Innovative Technology and Exploring Engineering* 8, 9 (2021), 853–860. https://www.irjet.net/archives/V8/i9/IRJET-V8I9127.pdf

[89] Jing Wen Taylor, Chuan Ching-Hua, Anghelcev George, Sar Sela, T. Yun Joseph, and Xu Yanzhen. 2024. Infusing affective computing models into advertising research on emotions. *Journal of Advertising* 53, 5 (2024), 710–731. https://doi.org/10.1080/00913367.2024.2409254 arXiv:https://doi.org/10.1080/00913367.2024.2409254

[90] Nattapong Thammasan, Ken-ichi Fukui, and Masayuki Numao. 2017. Multimodal fusion of EEG and musical features in music-emotion recognition. In *AAAI Conference on Artificial Intelligence*. AAAI Press, San Francisco, California, USA, 4991–4992. https://doi.org/10.1609/aaai.v31i1.11112

[91] Ha Thi Phuong Thao, Gemma Roig, and Dorien Herremans. 2023. EmoMV: Affective music-video correspondence learning datasets for classification and retrieval. *Information Fusion* 91 (2023), 64–79. https://doi.org/10.1016/j.inffus.2022.10.002

[92] Robert Thayer. 1989. *The Biopsychology of Mood and Arousal*. Oxford Academic, New York, NY, USA. https://doi.org/10.1093/oso/9780195068276.001.0001

[93] William F. Thompson. 2013. Intervals and scales. In *The Psychology of Music* (3rd ed.). Academic Press, London, UK, 107–140. https://doi.org/10.1016/B978-0-12-381460-9.00004-3

[94] William F. Thompson, E. Glenn Schellenberg, and Gabriela Husain. 2001. Arousal, mood, and the Mozart effect. *Psychological Science* 12, 3 (2001), 248–251. https://doi.org/10.1111/1467-9280.00345

[95] Guiying Tong. 2022. Multimodal music emotion recognition method based on the combination of knowledge distillation and transfer learning. *Scientific Programming* 2022, 1 (2022), 2802573. https://doi.org/10.1155/2022/2802573

[96] Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. 2008. Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech, and Language Processing* 16, 2 (2008), 467–476. https://doi.org/10.1109/TASL.2007.913750

[97] Jingyi Wang, Alireza Sharifi, Thippa Gadekallu, and Achyut Shankar. 2024. MMD-MII Model: A multilayered analysis and multimodal integration interaction approach revolutionizing music emotion classification. *International Journal of Computational Intelligence Systems* 17 (2024), 99. https://doi.org/10.1007/s44196-024-00489-6

[98] Shuo-Yang Wang, Ju-Chiang Wang, Yi-Hsuan Yang, and Hsin-Min Wang. 2014. Towards time-varying music auto-tagging based on CAL500 expansion. In *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, Chengdu, China, 1–6. https://doi.org/10.1109/ICME.2014.6890290

[99] Xing Wang, Chen Xiaoou, Deshun Yang, and Yuqian Wu. 2011. Music emotion classification of Chinese songs based on lyrics using TF*IDF and rhyme. In *International Society for Music Information Retrieval Conference (ISMIR)*. International Society for Music Information Retrieval, Miami, Florida, USA, 765–770. https://ismir2011.ismir.net/papers/PS6-19.pdf

[100] Yongjin Wang and Ling Guan. 2008. Recognizing human emotional state from audiovisual signals. *IEEE Transactions on Multimedia* 10, 5 (2008), 936–946. https://doi.org/10.1109/TMM.2008.927665

[101] Yao Wang, Zhu Liu, and Jin-Cheng Huang. 2000. Multimedia content analysis-using both audio and visual clues. *IEEE Signal Processing Magazine* 17, 6 (2000), 12–36. https://doi.org/10.1109/79.888862

[102] David Watson, David Wiese, Jatin Vaidya, and Auke Tellegen. 1999. The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence. *Journal of Personality and Social Psychology* 76, 5 (1999), 820–838. https://doi.org/10.1037/0022-3514.76.5.820

[103] Hanyu Xuan, Zhenyu Zhang, Shuo Chen, Jian Yang, and Yan Yan. 2020. Cross-modal attention network for temporal inconsistent audio-visual event localization. In *AAAI Conference on Artificial Intelligence*. AAAI Press, Palo Alto, California, USA, 279–286. https://doi.org/10.1609/aaai.v34i01.5361

[104] Hao Xue, Like Xue, and Feng Su. 2015. Multimodal music mood classification by fusion of audio and lyrics. In *International Conference on Multimedia Modeling (MMM)*. Springer International Publishing, Cham, Switzerland, 26–37. https://doi.org/10.1007/978-3-319-14442-9_3

[105] Liang Yang, Zhexu Shen, Jingjie Zeng, Xi Luo, and Hongfei Lin. 2024. COSMIC: Music emotion recognition combining structure analysis and modal interaction. *Multimedia Tools and Applications* 83, 5 (2024), 12519–12534. https://doi.org/10.1007/s11042-023-15376-z

[106] Qi Yang, Songhu Liu, and Tianzhuo Gong. 2025. Improve the application of reinforcement learning and multi-modal information in music sentiment analysis. *Expert Systems* 42, 1 (2025), e13416. https://doi.org/10.1111/exsy.13416

[107] Yi-Hsuan Yang and Homer H. Chen. 2012. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology* 3, 3 (2012), 40. https://doi.org/10.1145/2168752.2168754

[108] Yi-Hsuan Yang, Yu-Ching Lin, Heng-Tze Cheng, I-Bin Liao, Yeh-Chin Ho, and Homer H. Chen. 2008. Toward multi-modal music emotion classification. In *Pacific-Rim Conference on Multimedia (PCM)*. Springer, Berlin, Germany, 70–79. https://doi.org/10.1007/978-3-540-89796-5_8

[109] Guanghao Yin, Shouqian Sun, Dian Yu, Dejian Li, and Kejun Zhang. 2022. A multimodal framework for large-scale emotion recognition by fusing music and electrodermal activity signals. *ACM Transactions on Multimedia Computing, Communications, and Applications* 18, 3 (2022), 78. https://doi.org/10.1145/3490686

[110] Marcel Zentner, Didier Grandjean, and Klaus R Scherer. 2008. Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion* 8, 4 (2008), 494. https://doi.org/10.1037/1528-3542.8.4.494

[111] Kejun Zhang, Hui Zhang, Simeng Li, Changyuan Yang, and Lingyun Sun. 2018. The PMEmo dataset for music emotion recognition. In *International Conference on Multimedia Retrieval (ICMR)*. Association for Computing Machinery, New York, NY, USA, 135–142. https://doi.org/10.1145/3206025. 3206037

[112] Lige Zhang and Zhen Tian. 2022. Research on music emotional expression based on reinforcement learning and multimodal information. *Mobile Information Systems* 2022, 1 (2022), 2616220. https://doi.org/10.1155/2022/2616220

[113] Meixian Zhang, Yonghua Zhu, Wenjun Zhang, Yunwen Zhu, and Tianyu Feng. 2022. Modularized composite attention network for continuous music emotion recognition. *Multimedia Tools and Applications* 82, 5 (2022), 7319–7341. https://doi.org/10.1007/s11042-022-13577-6

[114] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, Wen Gao, and Qi Tian. 2018. Learning affective features with a hybrid deep model for audio-visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 10 (2018), 3030–3043. https://doi.org/10.1109/TCSVT.2017. 2719043

[115] Yong Zhang, Cheng Cheng, and Yidie Zhang. 2021. Multimodal emotion recognition using a hierarchical fusion convolutional neural network. *IEEE Access* 9 (2021), 7943–7951. https://doi.org/10.1109/ACCESS.2021.3049516

[116] Zongmeng Zhang, Xianjing Han, Xuemeng Song, Yan Yan, and Liqiang Nie. 2021. Multi-modal interaction graph convolutional network for temporal language localization in videos. *IEEE Transactions on Image Processing* 30 (2021), 8265–8277. https://doi.org/10.1109/TIP.2021.3113791

[117] Jiahao Zhao, Ganghui Ru, Yi Yu, Yulun Wu, Dichucheng Li, and Wei Li. 2022. Multimodal music emotion recognition with hierarchical cross-modal attention network. In *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, Taipei, Taiwan, 1–6. https://doi.org/10.1109/ICME52920. 2022.9859812

[118] Jiahao Zhao and Kazuyoshi Yoshii. 2023. Multimodal multifaceted music emotion recognition based on self-attentive fusion of psychology-inspired symbolic and acoustic features. In *Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, Taipei, Taiwan, 1641–1645. https://doi.org/10.1109/APSIPAASC58517.2023.10317539

[119] Jianchao Zhou, Xiaoou Chen, and Deshun Yang. 2019. Multimodel music emotion recognition using unsupervised deep neural networks. In *Conference on Sound and Music Technology (CSMT)*. Springer, Singapore, 27–39. https://doi.org/10.1007/978-981-13-8707-4_3