International Journal of Mathematical, Engineering and Management Sciences
Vol. 10, No. 4, 977-999, 2025
https://doi.org/10.33889/IJMEMS.2025.10.4.047

Ram Arti Publishers

Machine Learning and Deep Learning in Music Emotion Recognition: A Comprehensive Survey

Jumpi Dutta
Department of Electrical Engineering,
Assam Engineering College, Guwahati, Assam, India.
Corresponding Author: jumpiofcit@gmail.com

Dipankar Chanda
Department of Electrical Engineering,
Assam Engineering College, Guwahati, Assam, India.
E-mail: dchanda2007@rediffmail.com

(Received on August 19, 2024; Revised on November 6, 2024 & December 10, 2024 & January 27, 2025;
Accepted on February 2, 2025)

Abstract
Music can express and influence a wide range of emotional states and feelings in humans. The development of a system for recognizing emotions based on music analysis has generated significant interest among academic and industrial communities due to its applications in various fields such as human-machine interaction, music recommendation systems, music therapy, and so on. Music emotion recognition (MER) is the process of analysing and classifying the affective states conveyed by a piece of music. A survey of existing work on emotional music processing is indeed very helpful for carrying out further research in the field of music emotion recognition. Due to the importance of emotion recognition in Music Information Retrieval (MIR) research, a comprehensive survey is provided in this paper with a detailed study of emotion models, features, and various music databases. This paper emphasizes the machine learning and deep learning approaches used for MER to extract emotions from music. Finally, the paper is summarized with some possible future research directions.

Keywords- Music emotion recognition, Feature extraction, Machine learning, Deep learning, Human-computer-interaction, Music information retrieval.

1. Introduction

Music Information Retrieval (MIR) (Kaminskas & Ricci, 2012; Raieli, 2013) is an interdisciplinary field of science that deals with the development of various techniques to extract, analyse, and retrieve information from music (Music Information Retrieval, 2020). In recent years, the rapid development and expansion of digital music obtained from various sources have led to the expansion of digital music libraries, whereas conventional approaches are no longer adequate. Therefore, the rising demand for effective information access has necessitated the evolution of how music information is organized and retrieved (Fu et al., 2010; Masood et al., 2016). Some notable developments in music information retrieval (MIR) systems are content-based music retrieval, music similarity and clustering, artist identification (Fu et al., 2010; Tsai et al., 2017), genre classification (Chen & Wu, 2022; Fu et al., 2010, cover song detection (Fu et al., 2010; Yu et al., 2020), music emotion recognition (Cui et al., 2022; Dutta & Chanda, 2021; Louro et al., 2024a; Fu et al., 2010; Yang, 2021; Chen et al., 2015) etc.

Music emotions can be expressed and felt by listeners. Music serves as a suitable medium for emotional expression to convey feelings such as happiness, sadness, calm, and anger through musical elements including melody, harmony, rhythm, tempo, etc. This work mainly focuses on the emotions perceived by listeners, to understand how a listener interprets and recognizes the emotional content carried by music.

Music emotion recognition (MER) is an emerging and important sub-area of music information retrieval (MIR) that focuses on determining the emotional content of music using machine learning and deep learning techniques. Training of MER is performed using machine learning and deep learning techniques to build a mapping between the acoustic features of music and the perceived emotions, as annotated by listeners. It learns to predict the emotional content of music depending on its audio characteristics (Huq et al., 2010). The efficiency of a MER model can be estimated by comparing the emotions predicted by the model to the emotions perceived by listeners. If the model's predicted emotions closely align with the listener's emotions, then the model is considered accurate and effective in recognizing music emotions. The deep learning approach is mostly used for MER and natural language processing (Bahdanau et al., 2014; Khurana et al., 2022) as the deep learning models, particularly neural networks can analyse the data and perform the task effectively as compared to traditional feature extraction methods (Sarkar et al., 2019).

The tremendous emotional impact of music on human experiences and its vast range of applications motivates music emotion recognition (MER) research. Understanding emotions in music can improve music recommendation systems (Cheng et al., 2017; Verma et al., 2021), personalized music therapy (Sanyal et al., 2016; Sarkar et al., 2019), enhance user experiences in gaming and virtual reality, and support mental health interventions (Modran et al., 2023; Vuijk et al., 2023) through music therapy. Multimodal emotion recognition (Ahmed et al., 2023; Li et al., 2021) can help music therapists improve the emotional, physical, and social needs of individuals. Advances in MER provide better understanding of many cultures and personal preferences about music, therefore promoting inclusive music technology. The challenge of MER lies in identifying the emotions that music evokes, which can be complex due to the subjective nature of emotions in music itself. Music emotions are mainly influenced by the listener's experiences, cultural background, and contexts and hence different listeners may perceive the same music in varied ways.

This survey is necessary because existing research in MER has some limitations that prevent its efficacy and real-world applications. Most of the current research is focusing on western music, leading to insufficient representation of other cultural contexts. Due to a lack of multimodal and diverse datasets, MER research becomes challenging. A systematic framework for MER studies is also beneficial for the development in the field of research. Recognizing these gaps is important because it leads to more accurate and effective MER systems that can benefit multiple areas. Improved models for emotion recognition can result in more accurate music choices, better music therapy for mental health, and emotionally responsive experiences in gaming and virtual reality. It also helps reduce regional differences in how people feel emotion and helps to understand how musical structure and emotions affect each other.

The goal of the survey is: (1) to give detailed knowledge about the emotion model, different databases in MER, and relevant machine learning and deep learning approaches in the field of MER, (2) to put forward the issues faced in MER research and to unveil some possible directions for future research in the field of MER.

The paper is organized as follows: In Section 2, we discuss the research background of MER including emotion models, databases used in the field, features, and evaluation metrics. Machine learning and deep learning approaches in MER are introduced in Section 3. Section 4 focuses on issues related to MER along with development trends and some future directions in the field. Finally, Section 5 concludes the paper by providing a summary.

2. Background

Figure 1 shows the MER framework. In the preliminary stage, the samples in the music databases are pre-processed and relevant features are extracted from the samples. The samples are divided as the training and testing samples in the next step. Then suitable machine learning algorithm is applied to the extracted audio features to predict the emotions in the song samples.

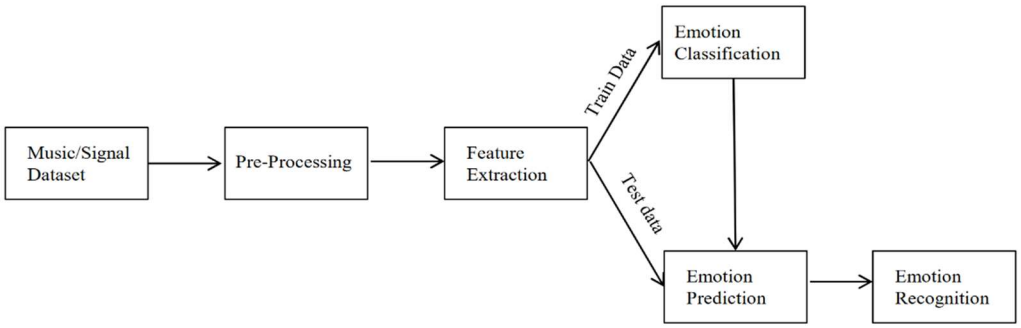

Figure 1. MER framework.

2.1 Emotion Models

Depending on the specific research goals, most of the accepted methods in the domain of MER are considered categorical and dimensional approaches (Griffiths et al., 2021). The categorical approach refers to a set of emotion tags assigned to a specific category that reflects its nature. The dimensional approach describes emotions along continuous dimensions or axes, referred to as affective dimensions (Patra et al., 2018). This approach considers emotions as points within a multidimensional space, instead of categorizing emotions into discrete labels. In **Table 1**, the most used emotion models in music are listed.

Table 1. Emotion models.

Emotion model	Domain	Emotion conceptualization	No of dimensions/classes
Hevner affective ring (Hevner, 1936; Yang, 2021)	Music	Categorical	67
Russell's circumplex model of affect (Russell, 1980)	General	Dimensional	2
Thayer model (Thayer, 1978)	General	Dimensional	2
GEMS (Chełkowska-Zacharewicz & Janowski, 2020; Zentner et al., 2008)	Music	Categorical	45
Larsen and Diener (1992)	General	Dimensional	2
PANA (Watson & Tellegen, 1985)	General	Dimensional	2
TWC model (Tellegen et al., 1999)	General	Dimensional	2

Hevner's effective ring model (Hevner, 1936; Yang, 2021) is indeed the earliest and most famous model in the field of music psychology and emotion research shown in **Figure 2**. The Hevner model comprises of 8 groups of 67 emotional adjectives arranged in a circle. Those eight groups of emotions are namely chant, sad, desire, lyrical, bouncing, happy, passionate, and solemnly (Yang, 2021).

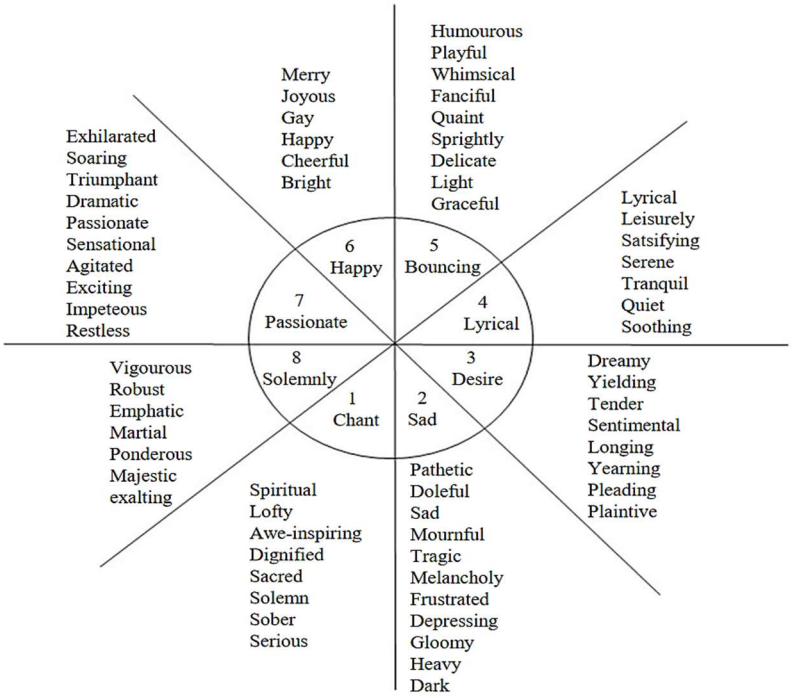

Figure 2. Hevner emotion model (Hevner, 1936).

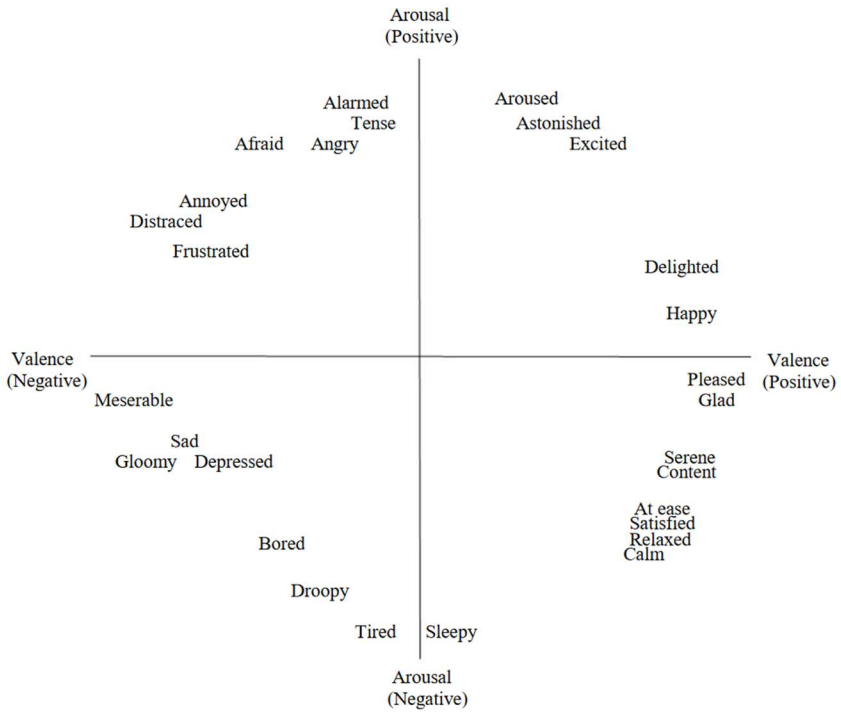

Figure 3. Russell emotion model (Russell, 1980).

In 1980, Russell (1980) introduced a circumplex model of effects in a two-dimensional space, with valence and arousal as the primary dimensions shown in **Figure 3**. This emotion model is one of the most popular and commonly used models in the field of emotion and psychology research. The positive (pleasant) and negative (unpleasant) emotions associated with a particular stimulus are represented as valence. Emotions with positive valence consist of happy, pleased, delighted, relaxed, etc., whereas emotions with negative valence consist of sad, angry, frustrated, bored, etc. Arousal represents the intensity level of an emotion. Emotions with low arousal (passive) are associated with contentment, tranquil, tired, etc, whereas emotions with high arousal (activated) consist of excited, astonished, tense etc (Jamdar et al., 2015).

Hevner's affective ring model (Hevner, 1936) and the Russell circumplex model of affects (Russell, 1980) are both different concepts, but both models have significant contributions in the field of music psychology and emotion research. Russell's model is considered to be the general model of affection, while Hevner's model mainly focuses on emotion in music.

Thayer's (1978) two-factor model of mood and emotion was proposed in 1978. Thayer two-dimensional model includes energetic arousal and tense arousal as its two dimensions, but the model doesn't include valence as a dimension just as in Russell's (1980) model. Energetic arousal reflects the level of mental energy adopted by an individual and tense arousal represents the psychological or physical tension adopted by an individual.

One of the popular emotion models developed for music-induced emotion is the Geneva Emotional Music Scales (GEMS) (Chełkowska-Zacharewicz & Janowski, 2020; Zentner et al., 2008). It consists of 45 emotional tags that are divided into nine emotional categories viz. wonder, transcendence, tenderness, nostalgia, peacefulness, power, joyful activation, tension, and sadness. All these emotions are considered first-order factors. The second-order factors are mainly a combination of the first-order factors namely sublimity, vitality, and unease. Sublimity is the combinations of the first-order factors namely wonder transcendence, tenderness, nostalgia, and peacefulness. Power and joyful activation combine to form vitality. Similarly, unease is the combination of tension and sadness.

Larsen and Diener (1992) proposed a new model defining the problem with the circumplex model of emotion in 1992. They proposed the two-dimensional model with pleasant-unpleasant in the horizontal axis and high activation-low activation in the vertical axis (Russell & Barrett, 1999). They observed that the circumplex model of emotion does not give a complete theory of emotion as the model consists of a wide range of emotions. Sometimes some of the emotions may not fit clearly in the predefined category. This may create confusion in the minds of researchers and different studies related to emotion detection.

Watson and Tellegen (1985) proposed a two-factor model of mood named as PANA (positive affect-negative affect) model (Rubin & Talarico, 2009). In their study, they rotated the Russell (Russell, 1980) circumplex model of emotion by 45 degrees to produce a new two-dimensional model of emotion representing the positive affect (PA) on the vertical axis and negative affect (NA) on the horizontal axis. According to the author, they found that PA and NA are the only dimensions that can describe a model of true emotions.

Tellegen et al. (1999) proposed a revised Thayer model of emotions referred to as the TWC (Tellegen-Watson-Clark) model. The TWC model expanded based on the two-dimensional (valence and arousal) Thayer model by incorporating 38 adjectives to describe emotions (Cui et al., 2022). This model introduced a new set of coordinates based on two primary emotional dimensions, i.e. happy and involved.

2.2 Emotional Music Databases

Database collection is one of the important and key tasks in a music emotion recognition system. Several criteria have to be kept in mind while considering a music database, such as the size of the database, the number of available emotions, appropriate annotation of music emotions, etc. Music emotions may be either perceived or induced. Induced emotions are felt by the listeners whereas perceived or expressed emotions are recognized by a listener as listening to the music (Song et al., 2016). Some of the pre-annotated music datasets like AMG1608 (Chen et al., 2015), and DEAM (Aljanaki et al., 2017; Goel, 2022) are available with emotional tags on each track. Some other datasets include different genres, and tempos.

Music emotional databases are developed to analyse the emotional content of music in different languages and across a wide range of emotional states. In MER, the attention is given to the listener's present state of mind (Lieskovská et al., 2021). However, due to copyright issues, some of the researchers work on the self-built database, while others work on available music database (Han et al., 2022; Yang & Chen, 2012). Different factors such as the singer's age, gender, and cultural background should be considered while creating self-built datasets. It is also essential to ensure high recording quality, sometimes in a soundproof room. Physiological tools such as EEG signals, heart rate, and skin conductance can sometimes be considered while preparing an emotional music dataset. Another issue with the music dataset is its limited size. Data augmentation (Pham et al., 2023) methods, such as time shifting, pitch shifting, etc can be employed to expand the dataset to compensate for this issue to some extent.

Developing a large and diverse database covering various music types, genres, and songs of different languages can be advantageous to make the database more general and inclusive. It can be observed that the emotion taxonomy of the existing work varies across different studies. The differences in emotion categories and datasets used can lead to variations in the various categories, such as emotion labels, emotion representation, model evaluation, etc. For better results, researchers can adopt the most widely accepted emotion taxonomies and also can create benchmark datasets that cover a wide range of emotions and domains (Yang & Chen, 2012). In **Table 2**, the most widely used databases in music are listed.

Table 2. A summary of music datasets.

Dataset name	Language	Emotion conceptualization	Emotions	No of songs /excerpts	Modalities	Perceived/ Induced
RAVDESS song database (Dutta & Chanda, 2024; Livingstone & Russo, 2018)	English	Categorical	Neutral, calm, happy, sad, angry, fearful	1012	Audio	Perceived
AMG1608 (Chen et al., 2015)	-	Dimensional	Valence and arousal	1608	Audio (WAV)	Perceived
AllMusic dataset (He & Ferguson, 2022; Panda et al., 2020)	English	Dimensional	Valence and arousal	900	-	
DEAM (Aljanaki et al., 2017; Goel, 2022)	English	Dimensional	Valence and arousal	1802	-	Perceived
Turkish emotional music database (TEM) (Hizlisoy et al., 2021)	Turkish	Dimensional	Valence and arousal	124 music excerpts	Audio	-
PMEmo dataset (He & Ferguson, 2022)	-	Dimensional	Valence and arousal	794	Audio (MP3)	Induced
MER500 (Goel, 2022)	Hindi	Categorical	Devotional, happy, party, romantic, sad	494	Audio	-
Emotify music database (Paolizzo et al., 2021)	-	Categorical	Classical, rock, pop and electronic music (genres)	400	Audio (MP3)	Induced
CAL500 (Barthet et al., 2013; Liu et al., 2017)	-	Categorical	174 labels	500	Audio (MP3)	Perceived

Table 2 continued…

DEAP dataset (Koelstra et al., 2012)	-	Dimensional	It records the EEG signal after listening to music, then checks the emotion	120 music video excerpts	-	Induced
AffectNet (Mollahosseini et al., 2019)	English, Spanish, Portuguese, German, Arabic, and Farsi	Both Categorical and Dimensional	Categorical: : Neutral, Happy, Sad, Surprise, Fear, Anger, Disgust, Contempt, None, Uncertain, and Non-face Dimensional: VA	1250	-	-
MER60 (Yang & Chen, 2011)	English	Dimensional	Valence and arousal	60	Audio	-
EmoMusic dataset (Bogdanov et al., 2022)	-	Dimensional	Valence and arousal	744	Audio	-
MusAV (Bogdanov et al., 2022)	-	Dimensional	Valence and arousal	2092	-	perceived
Soundtrack dataset (Han et al., 2023)	-	both categorical and dimensional	Categorical (tension, anger, fear, happy, sad, tender) and Dimensional (valence, energy, tension)	360 sound samples	Audio (MP3)	Perceived
Minnan songs datasets (Xiang et al., 2020)	Minnan	Categorical	Love, lovelorn, inspirational, lonely, homesickness, miss someone, and leave	1162	MP3	-
Hindi Song Dataset (Patra et al., 2018)	Hindi	-	5 mood classes: Excited, happy, calm, sad, and angry	1230 clips collected from 319 songs	MP3	-
BanglaMusicStylo (Mia et al., 2024)	Bengali	Categorical	Love, san and idealistic	2152	-	-
Assamese Song Database (Dutta & Chanda, 2024)	Assamese	Categorical	Calm, happy neutral and sad	200	Audio (WAV)	perceived

2.3 Features for Music Emotion Recognition

Feature extraction is one of the difficult aspects of MER as the quality of the feature extracted directly impacts on the system's ability to accurately categorize and recognize emotions in music (Han et al., 2022). An audio feature refers to the characteristics of an audio signal that may be extracted and analyzed to recognize the content of the sound signal. In the field of music analysis, audio features can be divided into low-level features, mid-level features, and high-level or top-level features (Fu et al., 2010).

2.3.1 Low-level Audio Features

Low-level audio features serve as the basic building blocks used in various audio processing tasks. These are the basic acoustic characteristics in music that are extracted directly from the audio signal. Low-level features can be divided into timbre (Chau et al., 2015) and temporal. Between these two features, timbre is considered as the short-term feature, while temporal is considered as the long-term feature (Fu et al., 2010). Timbre describes the quality or character of a sound that distinguishes it from other sounds (Yang et al., 2017) and is categorise as one of the commonly used low-level audio features. Different timbre features are the mel-frequency cepstrum coefficient (MFCC) (Huang et al., 2016; Yang & Chen, 2012), Daubechies wavelets coefficient histogram (DWCH) (Malheiro et al., 2016; Yang & Chen, 2012), linear predictive cepstrum coefficient (LPCC) (Rao & Nandi, 2015), zero crossing rate (ZCR) (Bhakre & Bang, 2016; Panda et al., 2023; Yang & Chen, 2011), Fourier cepstrum coefficient, spectral centroid (SC) (Ciaramella & Vettigli, 2013; Lerch, 2012), spectral rolloff (SR) (Lerch, 2012; Panda et al., 2023), spectral flux (SF) (Lerch, 2012; Panda et al., 2023), spectral bandwidth (SB), spectral crest factor (SCF) (Lerch, 2012; Panda et al., 2023), etc.

The temporal (Fu et al., 2010; Yang & Chen, 2012) feature captures the temporal evolution of a signal. Temporal features are constructed based on the patterns and changes observed in various aspects of the audio signal, including timbre features or spectrogram data. The popular and common type of temporal feature is the statistical moments namely mean variance, covariance, and kurtosis (Fu et al., 2010). In their work, Yang and Chen (2011) used temporal features like zero-crossing rate, temporal centroid, and long attack time to capture the temporal qualities of music.

Short-term features such as timbre capture the essential quality of the audio signal within small frames, typically around 10 to 100 milliseconds, while long-term features such as temporal consider the long-term effect, and are extracted from local windows with comparatively longer duration. Timbre and temporal features play different roles in audio processing. Timbre features are extracted from the smaller local windows; whereas temporal features come into play by considering the sequence or series of timbre features over longer texture windows (Fu et al., 2010).

2.3.2 Mid-level Audio Features

Mid-level features are derived from the top of low-level audio features. While low-level features have shown good performance for certain music classification tasks, they have some limitations in capturing the subjective aspects of music that humans find appealing. Mid-level features are mainly categorized as rhythm (Ciaramella & Vettigli, 2013; Schellenberg et al., 2000), pitch (Bhakre & Bang, 2016; Schellenberg et al., 2000), and harmony (Fu et al., 2010).

Rhythm is one of the widely used elements of music that represents the organization of time. It covers the pattern of long and short sounds, and silences observed in music (Panda et al., 2023). Rhythm is the patterns of pulses and notes with varying strengths that contribute to the overall structure of a musical piece (Yang & Chen, 2012). Lu et al. (2006) proposed five novel features or rhythmic properties of music namely rhythm strength, rhythm regularity, rhythm clarity, average tempo, and average onset frequency to express the three aspects (rhythm strength, rhythm regularity, and tempo) of rhythm that are closely related with people's mood. Rhythmic strength refers to the average onset strength where onset strength correlates with the amplitude or magnitude of the peak in the onset detection curve. Rhythm regularity means how consistent the rhythmic pattern is with the music. High rhythm regularity implies a consistent pattern, while low rhythm regularity implies irregular rhythm. Rhythm clarity relates to distinct and clear rhythmic features within the music. Average tempo is considered as the average speed or pace of the rhythm concerning bits per minute. A higher average tempo represents a faster rhythm. Average onset frequency may be obtained by dividing the total number of onsets by the time duration over which they occurred. A higher average onset frequency means a more rapid rhythm (Lu et al., 2006).

Pitch defines the perceived fundamental frequency of monophonic sound signals (Fu et al., 2010; Panda et al., 2023). High pitch value corresponds to emotions namely surprised, angry, fearful, happy, increased tense arousal, and low pitch value corresponds to sad, bored, pleasant, increased valence, etc. (Juslin & Laukka, 2004; Panda et al., 2023). Pitch class features, namely pitch class profile (PCP) (Gómez, 2006; Panda et al., 2023) and harmonic pitch class profile (HPCP) (Gómez, 2006; Panda et al., 2023) are essential tools for melody analysis and transcription tasks. The simplified version of HPCP is the chroma feature (Fu et al., 2010). The chroma feature represents the twelve different pitch classes and is one of the most powerful representations of music (Dutta & Chanda, 2021).

Harmony refers to the vertical aspects of music that involve the mixing of two or more notes played together to create chords (Fu et al., 2010; Panda et al., 2023). A chord is the fundamental building block of harmony. Harmony contributes to the overall tonal structure and emotion of a composition. Major

modes in harmony are mostly related to positive emotions, while others are related to negative emotions. Panda et al. (2023) discuss the effectiveness of harmony as one of the musical dimensions and its impact on emotions in music analysis.

2.3.3 High or Top-level Audio Features

High or top-level audio features refer to complex representations of audio content that capture high-level musical or semantic information. These features provide information that reflects how humans interpret music. High-level audio features include genre, emotion or mood, instruments, style, etc. Genres categorize music into different styles namely, rock, pop, classical, jazz, hip-hop, etc. Emotion-related features define music as happy, sad, angry, fearful, etc. Instrument label refers to the types of instruments used in music, such as piano, violin, guitar, drum, etc (Fu et al., 2010). Style represents a specific musical characteristic related to a particular culture, region, period, etc. (eg. classical, folk, blues, electronic).

2.4 Evaluation Metrics

Evaluation metrics are the benchmarks for determining and understanding the music emotion recognition systems. To evaluate the performance of the MER model, evaluation metrics are required. Based on the evaluation metrics, we can estimate the effectiveness of a proposed MER model. Accuracy, precision, sensitivity or recall value, and, F-score classification metrics are normally evaluated as the performance criteria from a model. Er and Esin (2021) use accuracy, precision, sensitivity, and, F-score to evaluate the performance criteria. Accuracy, precision, recall, and, F-score are used by Dutta and Chanda (2021) in their work on MER in Assamese songs. Accuracy may be defined as the total correct predictions made by the model to that of the total number of predictions (Er & Esin, 2021; Xiang et al., 2020). The formula for accuracy is:

$$\text{Accuracy} = \frac{\text{True Negative+True Positive}}{\text{False Negative+False Positive+True Negative+True Positive}} \quad (1)$$

Precision measures how many of the positive predictions or samples made by the classifier were correct predictions (Er & Esin, 2021; Xiang et al., 2020). The formula for precision is:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positives+Fa} \quad \text{Positives}} \quad (2)$$

Sensitivity or recall evaluates the ability of a classifier to correctly identify all positive instances out of the actual positive instances (Er & Esin, 2021; Xiang et al., 2020). Sensitivity is also known as recall or the true positive rate. The formula for sensitivity is:

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positives+Fal} \quad \text{Negative}} \quad (3)$$

F-score or F-measure (Dutta & Chanda, 2024; Er & Esin, 2021; Hizlisoy et al., 2021; Patra et al., 2018) is the harmonic mean of precision and recall. It is useful when there is an imbalance between positive and negative classes in a classification problem. The formula F-score is:

$$\text{F-score} = \frac{2*(\text{Precision}* \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (4)$$

There are several evaluation metrics used in regression problems, out of which R2 and RMSE (Root Mean Square Error) are the most popular metrics (Han et al., 2022; Yang, 2021). R2 is also known as the coefficient of determination and it ranges from 0 to 1. The upper range 1 means that the model perfectly fits the data and the lower value 0 means that the model does not fit any of the variance in the data. In a regression model, RMSE is the average error between the predicted values and the true values. A higher

RMSE value indicates a high prediction error. A lower RMSE value indicates a low prediction error and, hence more accurate predictions of the model to the data.

3. Machine Learning and Deep Learning in Music Emotion Recognition

Machine learning approach in music emotion recognition can be classified into two categories namely MER based on traditional machine learning models and MER based on deep learning models. Both categories are again sub-classified as categorical MER, dimensional MER, and music emotion variation detection (MEVD) (Aljanaki et al., 2017; Malik et al., 2017). The categorical approach to music emotion recognition involves categorizing emotions in discrete classes and then utilizing machine learning techniques to train a classifier (Yang & Chen, 2012). The dimensional approach to music emotion recognition represents emotions as numerical values within specific emotion dimensions, namely the valence and arousal dimensions. Valence represents the positivity (pleasant) and negativity (unpleasant) of an emotion. Arousal refers to the level of activation or intensity of an emotion (Russell, 1980). Music emotion variation detection (MEVD) captures the dynamic changes of music emotion, and MEVD involves making emotion predictions for short-time frames within a song, rather than predicting a single emotion label or value for an entire song. This gives a time series of emotion predictions by resulting more detailed understanding of how the emotional content of the music develops over time (Yang & Chen, 2012).

Another important approach to understanding emotions in music is static music emotion recognition (Louro et al., 2024b; Zhang et al., 2021). Static MER involves assigning a single, dominant emotion label to the whole music track. Due to a single emotion label, it is easier to implement static MER. Therefore, this approach can be used in music recommendation systems, playlist generation, etc. However, the major challenge with static MER is that it oversimplifies the emotion in a track by assigning a single label to the entire song. Many songs have evolved a range of emotions that vary by lyrics or section. Therefore, static MER does not capture the complex emotional experiences that music can offer. These limitations in static MER motivate the need for MEVD, which analyses music in segments and can label multi-emotional songs with different emotions.

There are some limitations of the categorical approach as it involves grouping musical expressions into a limited number of primary emotion classes. The range of primary emotion classes might be insufficient to express the diverse emotions that humans perceive in music. Categorizing the wide spectrum of musical emotions in a small set of discrete classes may not fully represent the diversity of musical emotions. Different individuals may interpret and label emotions in music in different ways as the language used to categorize emotions is ambiguous and can vary from one to another person. Moreover, having more emotion classes in studies related to MER makes it challenging to draw meaningful and appropriate conclusions from the study. Statistical analysis and interpreting the results from large emotion categories can be complex and time-consuming for researchers (Yang & Chen, 2012).

3.1 MER based on Machine Learning Models

MER based on machine learning models are summarized in **Table 3**.

3.1.1 Categorical MER

The most widely used classification models are support vector machine (SVM) (Chandrappa et al., 2023; Ciaramella & Vettigli, 2013; Er & Esin, 2021; Hizlisoy et al., 2021; Panda et al., 2015), decision tree (DT) (Salmam et al., 2016; Zuber & Vidhya, 2022), random forest (RF) (Hizlisoy et al., 2021; Gharsalli et al., 2016), k-nearest neighbors (KNN) (Hizlisoy et al., 2021; Jamdar et al., 2015; Panda et al., 2015), Multi-layer perceptron (MLP) (Ciaramella & Vettigli, 2013; Dutta & Chanda, 2021), Gaussian Mixture

Model (GMM) (Yang, 2021; Lu et al., 2006; Chen et al., 2017), Naïve Bayes (NB) (Bhakre & Bang, 2016; Sebe et al., 2002), etc. Dutta and Chanda (2021) achieved 93.75% accuracy using MFCC and MLP classifiers to recognize the emotions of Assamese music. Er and Esin (2021) worked on a self-built Turkish music dataset consisting of a total 400 number of samples with four classes in the dataset: happy, sad, angry, and relax. A total number of 34 feature values are obtained using MIRtoolbox and the data are classified using SVM, KNN, and ANN to evaluate the classification metrics. Finally, 79.30% accuracy is obtained using the ANN classifier using the normalized feature. Panda et al. (2015) performed their experiment of MER on standard and melodic audio features, and then input them into SVM, KNN, C4.5, and Naïve Bayes supervised algorithm for classification. SVM shows the best performance with 64% f-measure. They observed that high-level features are more relevant than low-level features in the MER study. Their model achieved poor results for the rhythm subset under standard audio using the SVM classifier. Paolizzo et al. (2021) work with an Emotify dataset that uses a version of emotion representation known as Geneva Emotion Music Scales-9 (GEMS-9) for automatic music emotion recognition. Ciaramella and Vettigli (2013) designed a framework that allows users to input a target song that represents their desired emotional state. Based on the user's input and preferences, the model generates a playlist of songs that match the emotional content specified by the user. However, they have performed the experiments only for 100 audio songs.

3.1.2 Dimensional MER

Yang et al. (2008) performed their experiment with various regression algorithms namely, support vector regression (SVR), multiple linear regression (MLR), and AdaBoost. RT (BoostR) (Solomatine & Shrestha, 2005), with a specific set of features as spectral contrast, DWCH (Daubechies wavelets coefficient histogram) algorithm (Yang & Chen, 2012), and features collected from PsySound (Cabrera, 1999) and MARSYAS (Yang & Chen, 2012; Tzanetakis & Cook, 2002). A total of 114-dimension features are extracted to train the three regression algorithms. Jamdar et al. (2015) proposed an approach to detect the emotion of a song based on its lyrical and audio features. Features like energy (Bhakre & Bang, 2016; Han et al., 2023), tempo, and danceability have been used with the k-Nearest Neighbors classifier for the emotion analysis of songs. They used a dimensional approach of songs namely valence and arousal and got an accuracy of 83.40% over 795 songs. Chen et al. (2017) use the AMG1608 dataset of 1608 30-second music clips. The ratings were collected from 665 listeners. The key finding of the proposed work is that modal adaptation with component tying significantly reduces the number of personal annotations required for effective personalization. It is reported that only 10-20 personal annotations achieve the same level of prediction accuracy as a baseline method that relies on 50 personal annotations. Yang (2021) improves the traditional back propagation (BP) neural network by introducing the artificial bee colony (ABC) algorithm. The output values generated by the ABC algorithm are used as the initial weights and threshold for the back propagation network. By incorporating the ABC algorithm, the BP neural network gains improved global search abilities. The influence of song genres on emotion recognition is observed by Koutras (2017). In the proposed system, determining the genre of a song can improve the recognition accuracy by a factor of 10-15%.

3.1.3 MEVD (Music Emotion Variation Detection)

Intensity, timbre, and, rhythm features are extracted by Lu et al. (2006) from a database comprising 800 music clips and then, fed to a GMM classifier to detect mood in audio clips. Four clusters namely Contentment, Depression, Exuberance, and Anxious/Frantic are considered for mood detection and 86.3% average accuracy was achieved for acoustic recordings of classical music. Chang et al. (2010) developed a personalized music recommendation system that can identify individual emotional responses to music. A correlation coefficient-based approach is applied to establish a relationship between the extracted audio feature and the emotional responses of listeners. The extracted audio features are then used to train the

SVM classifier for individual subjects and high classification accuracy is obtained by the proposed model. The work by Schmidt and Kim (2010) focuses on modelling and tracking emotions in the arousal-valence space using a two-dimensional Gaussian distribution model and Kalman filtering in a linear dynamical system to track how emotions change over time in response to audio and visual stimuli. However, they observed that the combination of MFCC and spectral contrasts works better with classification tasks but is unable to perform better with regression tasks.

Table 3. MER based on machine learning model.

Categories	Reference	Machine learning model	Emotion model	Dataset
Categorical MER	Dutta and Chanda (2021)	MLP classifier	Categorical	Self-built Assamese dataset
	Er and Esin (2021)	SVM, KNN, ANN	4-classes: happy, sad, angry, relax	Self-built Turkish music dataset
	Panda et al. (2015)	SVM, KNN, C4.5, and Naïve Bayes	Cluster 1 to 5	All Music database
	Paolizzo et al. (2021)	SMO-based SVM, ANN with back propagation, and Naïve Bayesian	9 emotions or classes: amazement, solemnity, tenderness, nostalgia, calmness, power, joyful, tension, sadness	Emotify music dataset
	Ciaramella and Vettigli (2013)	MLP, SVM and BN	Angry, Happy, Relax, Sad.	Self-built dataset of 100 audio songs
Dimensional MER	Yang et al. (2008)	SVR, MLR	VA model	Self-built database made up of 195 songs of Western, Chinese, and Japanese albums
	Jamdar et al. (2015)	KNN	VA model	Self-built music dataset prepared using last.fm social musical website
	Chen et al. (2017)	GMM	VA model	AMG1608 dataset
	Yang (2021)	SVM, KNN, GMM, BP, ABC algorithm	VA model (MAE, RMSE, R^2)	MediaEval Emotion in Music (MEM)
	Koutras (2017)	SVM	VA model (happy, angry, sad, peaceful)	Self-built dataset taken from Free Music Archive (FMA)consisting of 1100 song excerpts
MEVD	Lu et al. (2006)	GMM	Categorical	Self-built dataset
	Chang et al. (2010)	SVM	VA model	Self-built dataset
	Schmidt and Kim (2010)	multiple linear regression (MLR), a mixture of MLR and linear dynamical system (LDS)	VA model	Self-built dataset (consists of 15-second music clips from 240 songs)
	Yang et al. (2006)	Fuzzy KNN (FKNN), fuzzy nearest-mean classifier (FNM)	VA model	Self-built dataset collected from 243 songs from Western, Chinese, and Japanese language
	Xianyu et al. (2016)	DS-SVR	VA model	MediaEval emotion in music
	Schmidt et al. (2010)	SVM, SVR	VA model	Self-built dataset consisting of 15-second music clips from 240 songs.

Yang et al. (2006) observed the difficulty of assigning a single emotion class to a song segment in a deterministic manner. Two fuzzy classifiers are used in the work that measure the strength of the emotion related to a particular segment of music. This approach is useful for tracking how the emotions in a piece of music change over time. Xianyu et al. (2016) proposed a method with DS-SVR (Dual-stage Support Vector Regression), which is used for analyzing music and identifying mood changes within a song. DS-SVR employs two separate SVR models, one of them identifies mood changes between different songs, and the other model detects mood changes within a single song. Then both the results are combined to produce a final output. Schmidt et al. (2010) performed their experiment with time-varying features like MFCC, chroma, spectral shape, and spectral contrast with SVM and SVR classification methods. From

the MoodSwings database, 240 songs were collected with 15-seconds duration, and finally, they observed that MFCC and spectral contrast features give better results with the SVM and SVR classifiers.

3.2 MER based on Deep Learning Models

MER based on deep learning models are summarized in **Table 4**.

3.2.1 Categorical MER

Dutta and Chanda (2024) proposed a hybrid deep-learning algorithm using a CNN+LSTM classifier. RAVDESS emotional song database with six emotions and a self-built Assamese dataset with four emotions are used in their work. Relevant features namely MFCC, mel, and chroma are extracted from the given data and finally, they achieved the emotion recognition accuracy of 89.66% for the RAVDESS dataset and 85% for the Assamese dataset. Sarkar et al. (2019) proposed a novel technique using a convolutional neural network built around VGGNet (Visual Geometry Group Net) for MER. Three different datasets namely, Soundtracks, Bi-Modal (Malheiro et al., 2016), and MER_taffc (Panda et al., 2020) are used with suitable sets of features like spectral features, linear predictive coding, and MFCC. Their proposed model of CNN is the modified version of VGGNet with a smaller number of layers. Finally, they find that the model shows improved recognition accuracy by comparing the performance with three different systems. However, their model needs to be improved for low arousal values because significant confusion arises between sad and tender emotions as they belong to the low arousal Russell's plane. Agarwal et al. (2023) proposed deep-learning-based CNN and LSTM-GRU models for the emotion classification of musical data. Three deep-learning models were developed with different audio features, out of which the third model using mel-spectrographs and 2D CNN-LSTM model shows better performance in terms of F1 score and recall value as compared with existing methods. Pandeya et al. (2021) describe a multimodal research approach that influences the emotions carried out through music audio, video, and facial expressions. The model was tested across different unimodal and multimodal networks and was found to achieve high accuracy, F1 score, and AUC score.

3.2.2 Dimensional MER

Grekow (2021) uses an RNN network with various features to extract arousal and valence value. Two-layer RNNs performed better than the one-layer network with an increased number of LSTM units. Xia and Xu (2022) proposed a hybrid classifier model based on Thayer's two-dimensional emotion plane to achieve an accuracy of 84.9%. Polynomial regression, support vector regression, and k-plane piecewise regression are used in the training section and, then the input data is regressed and predicted to obtain its VA value in the test part. 3-4% improved accuracy is observed by the combined method of support vector regression and k-plane piecewise regression as compared to using one algorithm alone. A deep learning method of one-dimensional residual convolutional neural network (1D CNN) with the inception gate recurrent unit (GRU) was proposed by Han et al. (2023). They experimented on the Soundtrack dataset and observed that the proposed model performs effectively in emotion detection and classification tasks in music with an accuracy of 84%. An optimized model called brain emotional learning (BEL) combined with Thayer's psychological model was developed by Jandaghian et al. (2023) for music emotion recognition. The model comprises 12 emotional parts working simultaneously, with each part dedicated to evaluating a particular emotion based on Thayer's model. The proposed work offers a promising approach to performing across a diverse range of musical genres and reduces the limitations of existing MER studies. Aljanaki et al. (2017) developed a new benchmark for emotional analysis in music by introducing a new dataset MediaEval Database for Emotional Analysis in Music (DEAM), which is considered to be one of the largest datasets of dynamic annotation consisting of 1802 excerpts. They observed the recurrent neural networks (RNN) and mainly LSTM to be very powerful in expressing the dynamic changes in emotion in music from acoustic features. Hizlisoy et al. (2021) proposed a model of

music emotion recognition using low-level features related to timbre and energy, and a hybrid long-short-term memory + convolutional neural network (LSTM+CNN) classifier. They compared the performance of this classifier and observed improvements in music emotion recognition accuracies compared to k-nearest neighbor (k-NN), support vector machine (SVM), and random forest classifiers. They have achieved an accuracy of 99.19% using a new Turkish emotional music (TEM) database with duration of 30 seconds each. The model proposed by He and Ferguson (2022) uses segments as model input as it provides a suitable granularity for model training. The work presents a two-stage model that combines unsupervised and supervised learning algorithms for better training and performance compared to state-of-the-art models.

3.2.3 MEVD (Music Emotion Variation Detection)

Li et al. (2016) proposed a DBLSTM model that can capture information from both past and future contexts. DBLSTM is used to extract temporal context and hierarchical structure information in both directions. Post-processing is used to leverage the temporal correlation in the data and the fusion part of the model combines the result from DBLSTM with multiple times scales into a single prediction.

Table 4. MER based on a deep learning model.

Categories	Reference	Deep learning model	Emotion model	Dataset
Categorical MER	Dutta and Chanda (2024)	CNN+LSTM	6-emotion classes are: happy, sad, angry, calm, fearful, neutral 4-classes of emotions: calm, happy, neutral, sad	RAVDESS song database Self-built Assamese song dataset
	Sarkar et al. (2019)	CNN (VGGNet)	4-classes of emotions namely happy, anger, sad, and tender	Soundtracks, Bi-Modal, and MER traffc
	Agarwal et al. (2023)	CNN and LSTM-GRU model	happy, sad, romantic, dramatic, and aggressive	Kaggle open source dataset consists of 10133 samples
	Pandeya et al. (2021)	2D/3D CNN	6-emotion classes are: excited, fear, neutral, relaxation, sad and tension	Self-built music video dataset (data type: music, video and face)
Dimensional MER	Grekow (2021)	RNN	VA model	Publicly available GTZAN data
	Xia and Xu (2022)	polynomial regression, support vector regression, and k-plane piecewise regression	VA model	MediaEval
	Han et al. (2023)	1D CNN	4-classs of emotions namely happy, and anger (high arousal), sad and, tender (low arousal)	Soundtracks
	Jandaghian et al. (2023)	brain emotional learning (BEL) modified model	VA model	Set of music from Persian classical music (238 music samples)
	Aljanaki et al. (2017)	LSTM-RNN	VA model	DEAM
	Hizlisoy et al. (2021)	LSTM+CNN	VA model	Turkish emotional music (TEM) database
	He and Ferguson (2022)	Bidirectional-LSTM (BiLSTM)	VA model	PMEmo and AllMusic
MEVD	Li et al. (2016)	DBLSTM	VA model	MediaEval emotion in music
	Jia (2022)	CNN-LSTM and explicit sparse attention network	4 emotions namely happy, sad, relax, and anger	2147 Chinese music samples

The experiment was carried out with the MediaEval 2015 dataset and it is found that DBLSTM alone is more effective in predicting emotions as compared to MLR or SVR. Jia (2022) proposed a music emotion recognition and classification model using CNN-LSTM and explicit sparse attention network to improve

the model accuracy of existing work. The experiment was carried out with a Chinese complex music dataset consisting of 2147 samples and four different emotions. An explicit sparse attention network is a mechanism of deep learning models that focuses on relevant features which reduces the effect of irrelevant information and thereby improves the efficiency of the model. The model shows a recognition accuracy of 0.71 for happy, 0.688 for sad, 0.659 for relaxed, 0.651 for angry, and 0.677 for average accuracy.

4. Discussion

4.1 Issues Related to MER

Some of the important issues related to music emotion recognition are discussed below in brief.

- *Lack of data*: The availability of standard, large-scale, diverse emotion-labeled music datasets is crucial for advancing MER research. There is a limitation in the existing dataset in terms of its limited size and genre diversity, which can affect the development of MER systems applicable to a wide range of music genres. Researchers often create their datasets, which can vary in terms of quality, size, and diversity and thus it becomes difficult to extract meaningful conclusions or compare findings across different studies.

- *Data annotation*: Dimensional MER systems need numerical emotion ratings, which are sometimes not readily available in online repositories. It may be challenging to provide precise numerical ratings for emotions on a continuous scale. Moreover, ensuring consistency in emotion rating scales across different subjects or within the same subject can be difficult. Finally, inconsistent and imprecise emotion ratings can lead to inaccuracies in the training and evaluation of MER models. Ranking-based emotion annotations can reduce this problem to some extent. It is necessary to develop more reliable and standardized scales for dimensional emotion ratings to improve the overall accuracy of MER models.

- *Lack of consensus on affective terms* (*Emotional labeling of music*): Different people may perceive and interpret emotions in music differently. There is a lack of consensus among different people regarding which affective terms are best suited for the emotional content of a music piece. Therefore, traditional categorical approaches may not perform well in practice as they aim to assign only one emotion class to each music piece. Even in dimensional approaches, different listeners can have varying emotional responses to the same song, leading to variations in dimensional emotion ratings. The emotional labeling of music excerpts is one of the main challenges in MER. This implies determining the emotional categories that are best suitable for a particular piece of music. Appropriate labeling of music is necessary for training machine learning models.

- *Problem with mixed or compound emotions*: Some music pieces may contain mixed or compound emotions such as sadly surprised, sadly disgusted, happily surprised, happily disgusted, angrily fearful, etc. Therefore, researchers and developers working on MER systems may need to implement strategies or techniques to deal with such types of compound emotions. A continuous representation in a complex emotion space can be used to capture the nuances of such mixed emotions.

- *Limitation of conventional methods*: Low-level features may have limitations in capturing the emotional content in music. Conventional MER models primarily rely on low-level audio features to predict emotions. To improve MER performance, it is necessary to select factors beyond low-level audio features. The selection of high-level features such as harmony, melody, and rhythm in the MER

system may improve the performance of MER. Also, integrating information from multiple modalities, such as lyrics, audio, and music theory, can create more effective emotional content for music models.

4.2 Development Trends and Future Research Directions

A comprehensive survey of the existing research in the field of MER provides a strong foundation for future research. It helps to contribute to the advancement of the field by addressing current challenges and limitations. The field of MER has tremendous research potential in the last few years. The recent survey suggested the following development status and future research directions.

- The use of multimodal emotion recognition in MER can significantly enhance the accuracy and depth of understanding of a person's emotional response to music. Facial expression, lyrics of songs, brain signal (eg. EEG), and physiological signals can be employed to assess the emotional response of a person while listening to music. EEG is a powerful tool to study brain activity. The brain activity of a person can be detected while listening to music and then machine learning can help to decode this pattern to recognize emotion. An available database known as the DEAP database (Koelstra et al., 2012) uses the EEG signal. Other physiological signals like heart rate, body temperature, and skin conductance can also be used to detect emotion while listening to music. The changes in these parameters can give a measurement of emotional arousal and valence. Multimodal emotion recognition can help music therapists to improve the emotional, physical, and social needs of individuals.

- Interdisciplinary collaboration: The collaboration with experts from psychology and musicology may help the researchers to better understand the correlation between emotional responses and musical features. The interdisciplinary approach has led to more extensive research.

The lack of a standard and high-quality emotional dataset is a challenge in MER tasks. In 2007, Music Information Retrieval Evaluation eXchange (MIREX), a prestigious international audio retrieval and evaluation competition, included audio mood classification (AMC) as one of its tasks. This competition emphasizes the increasing significance of MER (Han et al., 2022). The MIREX mood classification dataset is one of the efforts to solve the problem of standard music databases. While MIREX datasets are valuable resources for certain tasks, they may not be publicly available or suitable for some of the emotion recognition research (Panda et al., 2015).

5. Conclusion

Music emotion recognition is a fascinating and fast-growing research field that explores the use of machine learning and artificial intelligence to analyze the emotional content of music. The MER research has developed rapidly in the last few decades. With the advancement of machine learning, deep learning, and neural networks, the capabilities of recognizing music have grown significantly. However, challenges are still there as emotions are subjective. Due to several factors, including individual experiences, cultural backgrounds, and psychological factors, a person's experiences of emotion in music can vary widely. This systematic survey can be helpful in areas like music recommendation systems, personal wellness, music therapy, and advertising and marketing. Streaming platforms like Spotify and Apple Music can recommend music playlists based on the users' moods to enhance their present emotional state. Smart homes or workplaces may be designed to increase work productivity with MER-enabled smart devices. Therapists can use music with positive emotions to support conditions like anxiety and depression in a patient. Emotionally resonant music in marketing and advertisement fields can influence many customers with positive shopping experiences.

This survey has significantly improved the intended field of research by providing a systematic framework for future work in the field of MER. This work provides information on a variety of datasets in a number of languages, some of which have been introduced recently. Therefore, these datasets may be employed to conduct a significant amount of research. In addition, this work analyzed the emotion models, feature extraction, and classification techniques in MER using deep learning and machine learning techniques. Then, a discussion is conducted on the issues associated with MER research and its potential future directions. Thus, a summary of the latest trends, methodologies, and approaches in MER helps researchers in recognizing the latest advancements and identifying gaps in the existing literature.

Conflict of Interest

The authors confirm that there is no conflict of interest to declare for this publication.

AI Disclosure

The author(s) declare that no assistance is taken from generative AI to write this article.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Agarwal, G., Gupta, S., Agarwal, S., & Rai, A.K. (2023). Emotion classification for musical data using deep learning techniques. *International Journal of Reconfigurable and Embedded Systems, 12*(2), 240. https://doi.org/10.11591/ijres.v12.i2.pp240-247.

Ahmed, N., Aghbari, Z.A., & Girija, S. (2023). A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications*, *17*, 200171. https://doi.org/10.1016/j.iswa.2022.200171.

Aljanaki, A., Yang, Y.H., & Soleymani, M. (2017). Developing a benchmark for emotional analysis of music. *PLoS ONE*, *12*(3), e0173392. https://doi.org/10.1371/journal.pone.0173392.

Bahdanau, D. Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*. https://doi.org/10.48550/arxiv.1409.0473.

Barthet, M., Fazekas, G., & Sandler, M. (2013). Music emotion recognition: From content-to context-based models. In *From Sounds to Music and Emotions: 9th International Symposium* (pp. 228-252). Springer, Berlin, Heidelberg, London, UK.

Bhakre, S.K., & Bang, A. (2016). Emotion recognition on the basis of audio signal using Naive Bayes classifier. In *2016 International Conference on Advances in Computing, Communications and Informatics* (pp. 2363-2367). IEEE. Jaipur, India. https://doi.org/10.1109/icacci.2016.7732408.

Bogdanov, D., Lizarraga-Seijas, X., Alonso-Jiménez, P., & Serra, X. (2022). MusAV: a dataset of relative arousal-valence annotations for validation of audio models. *Proceedings of the 23rd International Society for Music Information Retrieval Conference* (pp. 05-15). Bengaluru, India. https://doi.org/10.5281/zenodo.7316746.

Cabrera, D. (1999). Psysound: A computer program for psychoacoustical analysis. In *Proceedings of the Australian Acoustical Society Conference* (Vol. 24, pp. 47-54). Melbourne, Australia.

Chandrappa, S., Shekar, P.C., Chaya, P., Dharmanna, L., & Guruprasad, M.S. (2023). Machine learning algorithms for identifying fake currencies. *SN Computer Science*, *4*(4), 368. https://doi.org/10.1007/s42979-023-01812-2.

Chang, C.Y., Lo, C.Y., Wang, C.J., & Chung, P.C. (2010). A music recommendation system with consideration of personal emotion. In *2010 International Computer Symposium* (pp. 18-23). IEEE. Tainan, Taiwan.

Chau, C., Wu, B., & Horner, A. (2015). The emotional characteristics and timbre of nonsustaining instrument sounds. *Journal of the Audio Engineering Society*, *63*(4), 228-244. https://doi.org/10.17743/jaes.2015.0016.

Chełkowska-Zacharewicz, M., & Janowski, M. (2020). Polish adaptation of the Geneva emotional music scale: factor structure and reliability. *Psychology of Music*, *49*(5), 1117-1131.

Chen, W., & Wu, G. (2022). A multimodal convolutional neural network model for the analysis of music genre on children's emotions influence intelligence. *Computational Intelligence and Neuroscience*, *2022*(1), 5611456.

Chen, Y.A., Wang, J.C., Yang, Y.H., & Chen, H.H. (2017). Component tying for mixture model adaptation in personalization of music emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(7), 1409-1420. https://doi.org/10.1109/taslp.2017.2693565.

Chen, Y.A., Yang, Y.H., Wang, J.C., & Chen, H. (2015). The AMG1608 dataset for music emotion recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 693-697). IEEE. South Brisbane, QLD, Australia. https://doi.org/10.1109/icassp.2015.7178058.

Cheng, Z., Shen, J., Zhu, L., Kankanhalli, M.S., & Nie, L. (2017). Exploiting music play sequence for music recommendation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (Vol. 17, pp. 3654-3660). AAAI press, Melbourne, Australia. https://doi.org/10.24963/ijcai.2017/511.

Ciaramella, A., & Vettigli, G. (2013). Machine learning and soft computing methodologies for music emotion recognition. In *Neural Nets and Surroundings: 22nd Italian Workshop on Neural Nets* (pp. 427-436). Springer, Berlin, Heidelberg, Salerno, Italy. https://doi.org/10.1007/978-3-642-35467-0_42.

Cui, X., Wu, Y., Wu, J., You, Z., Xiahou, J., & Ouyang, M. (2022). A review: music-emotion recognition and analysis based on EEG signals. *Frontiers in Neuroinformatics*, *16*, 997282. https://doi.org/10.3389/fninf.2022.997282.

Dutta, J., & Chanda, D. (2021). Music emotion recognition in assamese songs using MFCC features and MLP classifier. In *2021 International Conference on Intelligent Technologies* (pp. 1-5). IEEE. Hubli, India. https://doi.org/10.1109/conit51480.2021.9498345.

Dutta, J., & Chanda, D. (2024). Music emotion recognition and classification using hybrid CNN-LSTM deep neural network. *Bangladesh Journal of Multidisciplinary Scientific Research*, *9*(3), 21-32. https://doi.org/10.46281/bjmsr.v9i3.2230

Er, M.B., & Esin, E.M. (2021). Music emotion recognition with machine learning based on audio features. *Computer Science*, *6*(3), 133-144. https://doi.org/10.53070/bbd.945894.

Fu, Z., Lu, G., Ting, K.M., & Zhang, D. (2010). A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia*, *13*(2), 303-319. https://doi.org/10.1109/tmm.2010.2098858.

Gharsalli, S., Emile, B., Laurent, H., & Desquesnes, X. (2016). Feature selection for emotion recognition based on random forest. In *International Conference on Computer Vision Theory and Applications* (Vol. 5, pp. 610-617). SciTePress. https://doi.org/10.5220/0005725206100617.

Goel, S. (2022). Emotion classification using nature based optimization with transformers and transfer learning. *Journal of Pharmaceutical Negative Results*, *13*(10), 3052-3071. https://doi.org/10.47750/pnr.2022.13.S10.369.

Gómez, E. (2006). *Tonal description of music audio signals* [PhD thesis]. Universitat Pompeu Fabra.

Grekow, J. (2021). Music emotion recognition using recurrent neural networks and pretrained models. *Journal of Intelligent Information Systems*, *57*(3), 531-546. https://doi.org/10.1007/s10844-021-00658-5.

Griffiths, D., Cunningham, S., Weinel, J., & Picking, R. (2021). A multi-genre model for music emotion recognition using linear regressors. *Journal of New Music Research*, *50*(4), 355-372. https://doi.org/10.1080/09298215.2021.1977336.

Han, D., Kong, Y., Han, J., & Wang, G. (2022). A survey of music emotion recognition. *Frontiers of Computer Science*, *16*(6), 166335. https://doi.org/10.1007/s11704-021-0569-4.

Han, X., Chen, F., & Ban, J. (2023). Music emotion recognition based on a neural network with an inception-gru residual structure. *Electronics*, *12*(4), 978. https://doi.org/10.3390/electronics12040978.

He, N., & Ferguson, S. (2022). Music emotion recognition based on segment-level two-stage learning. *International Journal of Multimedia Information Retrieval*, *11*(3), 383-394. https://doi.org/10.1007/s13735-022-00230-z.

Hevner, K. (1936). Experimental studies of the elements of expression in music. *The American Journal of Psychology*, *48*(2), 246. https://doi.org/10.2307/1415746.

Hizlisoy, S., Yildirim, S., & Tufekci, Z. (2021). Music emotion recognition using convolutional long short term memory deep neural networks. *Engineering Science and Technology an International Journal*, *24*(3), 760-767. https://doi.org/10.1016/j.jestch.2020.10.009.

Huang, Z., Xue, W., Mao, Q., & Zhan, Y. (2016). Unsupervised domain adaptation for speech emotion recognition using PCANet. *Multimedia Tools and Applications*, *76*(5), 6785-6799. https://doi.org/10.1007/s11042-016-3354-x.

Huq, A., Bello, J.P., & Rowe, R. (2010). Automated music emotion recognition: a systematic evaluation. *Journal of New Music Research*, *39*(3), 227-244. https://doi.org/10.1080/09298215.2010.513733.

Jamdar, A., Abraham, J., Khanna, K., & Dubey, R. (2015). Emotion analysis of songs based on lyrical and audio features. *International Journal of Artificial Intelligence & Applications*, *6*(3), 35-50. https://doi.org/10.5121/ijaia.2015.6304.

Jandaghian, M., Setayeshi, S., Razzazi, F., & Sharifi, A. (2023). Music emotion recognition based on a modified brain emotional learning model. *Multimedia Tools and Applications*, *82*(17), 26037-26061. https://doi.org/10.1007/s11042-023-14345-w.

Jia, X. (2022). Music emotion classification method based on deep learning and explicit sparse attention network. *Computational Intelligence and Neuroscience*, *2022*(1), 3920663. https://doi.org/10.1155/2022/3920663.

Juslin, P.N., & Laukka, P. (2004). Expression, perception, and induction of musical emotions: a review and a questionnaire study of everyday listening. *Journal of New Music Research*, *33*(3), 217-238. https://doi.org/10.1080/0929821042000317813.

Kaminskas, M., & Ricci, F. (2012). Contextual music information retrieval and recommendation: State of the art and challenges. *Computer Science Review*, *6*(2-3), 89-119. https://doi.org/10.1016/j.cosrev.2012.04.002.

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*(3), 3713-3744. https://doi.org/10.1007/s11042-022-13428-4.

Koelstra, S., Muhl, C., Soleymani, M., Lee, N.J., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., & Patras, I. (2012). DEAP: a database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, *3*(1), 18-31. https://doi.org/10.1109/t-affc.2011.15.

Koutras, A. (2017). Song emotion recognition using music genre information. In *Speech and Computer: 19th International Conference* (pp. 669-679). Springer International Publishing. Hatfield, UK. https://doi.org/10.1007/978-3-319-66429-3_67.

Larsen, R.J., & Diener, E. (1992). Promises and problems with the circumplex model of emotion. In Clark, M.S. (ed) *Review of Personality and Social Psychology, Emotion* (pp. 25-59). Sage Publications, Inc. Thousand Oaks, CA, US.

Lerch, A. (2012). *An introduction to audio content analysis*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Li, M., Qiu, X., Peng, S., Tang, L., Li, Q., Yang, W., & Ma, Y. (2021). Multimodal emotion recognition model based on a deep neural network with multiobjective optimization. *Wireless Communications and Mobile Computing*, *2021*(1), 6971100. https://doi.org/10.1155/2021/6971100.

Li, X., Tian, J., Xu, M., Ning, Y., & Cai, L. (2016). DBLSTM-based multi-scale fusion for dynamic emotion prediction in music. In *2016 IEEE International Conference on Multimedia and Expo* (pp. 1-6). IEEE. Seattle, WA, USA. https://doi.org/10.1109/icme.2016.7552956.

Lieskovská, E., Jakubec, M., Jarina, R., & Chmulík, M. (2021). A review on speech emotion recognition using deep learning and attention mechanism. *Electronics*, *10*(10), 1163. https://doi.org/10.3390/electronics10101163.

Liu, X., Chen, Q., Wu, X., Liu, Y., & Liu, Y. (2017). CNN based music emotion classification. *arXiv preprint arXiv:1704.05665*. https://doi.org/10.48550/arxiv.1704.05665.

Livingstone, S.R., & Russo, F.A. (2018). The Ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English. *PloS One*, *13*(5), e0196391. https://doi.org/10.1371/journal.pone.0196391.

Louro, P.L., Redinho, H., Malheiro, R., Paiva, R.P., & Panda, R. (2024a). A comparison study of deep learning methodologies for music emotion recognition. *Sensors*, *24*(7), 2201. https://doi.org/10.3390/s24072201.

Louro, P.L., Redinho, H., Santos, R., Malheiro, R., Panda, R., & Paiva, R.P. (2024b). MERGE--a bimodal dataset for static music emotion recognition. *arXiv preprint arXiv:2407.06060*. https://doi.org/10.48550/arxiv.2407.06060.

Lu, N.L., Liu, D., & Zhang, N.H. (2006). Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio Speech and Language Processing*, *14*(1), 5-18. https://doi.org/10.1109/tsa.2005.860344.

Malheiro, R., Panda, R., Gomes, P.J., & Paiva, R.P. (2016). Bi-modal music emotion recognition: Novel lyrical features and dataset. In *9th International Workshop on Music and Machine Learning–MML 2016–in Conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases.* Riva del Garda, Italy.

Malik, M., Adavanne, S., Drossos, K., Virtanen, T., Ticha, D., & Jarina, R. (2017). Stacked convolutional and recurrent neural networks for music emotion recognition. *arXiv preprint arXiv:1706.02292*.

Masood, S., Nayal, J.S., & Jain, R.K. (2016). Singer identification in Indian Hindi songs using MFCC and spectral features. In *2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems* (pp. 1-5). IEEE. Delhi, India. https://doi.org/10.1109/icpeices.2016.7853641.

Mia, M., Das, P., & Habib, A. (2024). Verse-based emotion analysis of Bengali music from lyrics using machine learning and neural network classifiers. *International Journal of Computing and Digital Systems*, *15*(1), 359-370. https://doi.org/10.12785/ijcds/150128.

Modran, H.A., Chamunorwa, T., Ursuțiu, D., Samoilă, C., & Hedeșiu, H. (2023). Using deep learning to recognize therapeutic effects of music based on emotions. *Sensors*, *23*(2), 986. https://doi.org/10.3390/s23020986.

Mollahosseini, A., Hasani, B., & Mahoor, M.H. (2019). AffectNet: a database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, *10*(1), 18-31. https://doi.org/10.1109/taffc.2017.2740923.

Music Information retrieval. (2020). In *En.wikipedia.org, 2020*. Retrieved August 16, 2024, from https://en.wikipedia.org/wiki/Music_information_retrieval.

Panda, R., Malheiro, R., & Paiva, R.P. (2020). Novel audio features for music emotion recognition. *IEEE Transactions on Affective Computing*, *11*(4), 614-626. https://doi.org/10.1109/taffc.2018.2820691.

Panda, R., Malheiro, R., & Paiva, R.P. (2023). Audio features for music emotion recognition: a survey. *IEEE Transactions on Affective Computing*, *14*(1), 68-88. https://doi.org/10.1109/taffc.2020.3032373.

Panda, R., Rocha, B., & Paiva, R.P. (2015). Music emotion recognition with standard and melodic audio features. *Applied Artificial Intelligence*, *29*(4), 313-334. https://doi.org/10.1080/08839514.2015.1016389.

Pandeya, Y.R., Bhattarai, B., & Lee, J. (2021). Deep-learning-based multimodal emotion classification for music videos. *Sensors*, *21*(14), 4927. https://doi.org/10.3390/s21144927.

Paolizzo, F., Pichierri, N., Giardino, D., Matta, M., Casali, D., & Costantini, G. (2021). A new multilabel system for automatic music emotion recognition. In *2021 IEEE International Workshop on Metrology for Industry 4.0 & IoT* (pp. 625-629). IEEE. Rome, Italy. https://doi.org/10.1109/metroind4.0iot51437.2021.9488537.

Patra, B.G., Das, D., & Bandyopadhyay, S. (2018). Multimodal mood classification of Hindi and Western songs. *Journal of Intelligent Information Systems*, *51*(3), 579-596. https://doi.org/10.1007/s10844-018-0497-4.

Pham, N.T., Dang, D.N.M., Nguyen, N.D., Nguyen, T.T., Nguyen, H., Manavalan, B., Lim, C.P., & Nguyen, S.D. (2023). Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition. *Expert Systems with Applications*, *230*, 120608. https://doi.org/10.1016/j.eswa.2023.120608.

Raieli, R. (2013). *Multimedia information retrieval: theory and techniques*. Chandos Publication, Philadelphia.

Rao, K.S., & Nandi, D. (2015). *Language identification using excitation source features*. Springer International Publishing. https://doi.org/10.1007/978-3-319-17725-0.

Rubin, D.C., & Talarico, J.M. (2009). A comparison of dimensional models of emotion: evidence from emotions, prototypical events, autobiographical memories, and words. *Memory*, *17*(8), 802-808. https://doi.org/10.1080/09658210903130764.

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

Russell, J.A., & Barrett, L.F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, *76*(5), 805-819. https://doi.org/10.1037/0022-3514.76.5.805.

Salmam, F.Z., Madani, A., & Kissi, M. (2016). Facial expression recognition using decision trees. In *2016 13th International Conference on Computer Graphics, Imaging and Visualization* (pp. 125-130). IEEE. Beni Mellal, Morocco. https://doi.org/10.1109/cgiv.2016.33.

Sanyal, S., Banerjee, A., Sengupta, R., & Ghosh, D. (2016). Chaotic brain, musical mind-a non-linear neurocognitive physics based study. *Journal of Neurology and Neuroscience*, *7*(63), 1-10. https://doi.org/10.21767/2171-6625.100063.

Sarkar, R., Choudhury, S., Dutta, S., Roy, A., & Saha, S.K. (2019). Recognition of emotion in music based on deep convolutional neural network. *Multimedia Tools and Applications*, *79*(1-2), 765-783. https://doi.org/10.1007/s11042-019-08192-x.

Schellenberg, E.G., Krysciak, A.M., & Campbell, R.J. (2000). Perceiving emotion in melody: interactive effects of pitch and rhythm. *Music Perception an Interdisciplinary Journal*, *18*(2), 155-171. https://doi.org/10.2307/40285907.

Schmidt, E.M., & Kim, Y.E. (2010). Prediction of time-varying musical mood distributions using Kalman filtering. In *2010 Ninth International Conference on Machine Learning and Applications* (pp. 655-660). IEEE. Washington, DC, USA. https://doi.org/10.1109/icmla.2010.101.

Schmidt, E.M., Turnbull, D., & Kim, Y.E. (2010). Feature selection for content-based, time-varying musical emotion regression. In *Proceedings of the International Conference on Multimedia Information Retrieval* (pp. 267-274). ACM. https://doi.org/10.1145/1743384.1743431.

Sebe, N., Lew, M.S., Cohen, I., Garg, A., & Huang, T.S. (2002). Emotion recognition using a Cauchy Naive Bayes classifier. In *2002 International Conference on Pattern Recognition* (Vol. 1, pp. 17-20). IEEE. Quebec City, Canada. https://doi.org/10.1109/icpr.2002.1044578.

Solomatine, D.P., & Shrestha, D.L. (2004). AdaBoost. RT: a boosting algorithm for regression problems. In *2004 IEEE International Joint Conference on Neural Networks* (Vol. 2, pp. 1163-1168). IEEE. Budapest, Hungary.

Song, Y., Dixon, S., Pearce, M.T., & Halpern, A.R. (2016). Perceived and induced emotion responses to popular music: categorical and dimensional models. *Music Perception: An Interdisciplinary Journal*, *33*(4), 472-492.

Tellegen, A., Watson, D., & Clark, L.A. (1999). On the dimensional and hierarchical structure of affect. *Psychological Science*, *10*(4), 297-303. https://doi.org/10.1111/1467-9280.00157.

Thayer, R.E. (1978). Toward a psychological theory of multidimensional activation (arousal). *Motivation and Emotion*, *2*(1), 1-34. https://doi.org/10.1007/bf00992729.

Tsai, T., Pratzlich, T., & Muller, M. (2017). Known-artist live song identification using audio hashprints. *IEEE Transactions on Multimedia*, *19*(7), 1569-1582. https://doi.org/10.1109/tmm.2017.2669864.

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, *10*(5), 293-302. https://doi.org/10.1109/tsa.2002.800560.

Verma, N.V., Marathe, N.N., Sanghavi, N.P., & Nitnaware, N.D.P. (2021). Music recommendation system using machine learning. *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, *7*(6), 80-88. https://doi.org/10.32628/cseit217615.

Vuijk, J.G.J., Brinke, J.K., & Sharma, N. (2023). Utilising emotion monitoring for developing music interventions for people with dementia: a state-of-the-art review. *Sensors*, *23*(13), 5834. https://doi.org/10.3390/s23135834.

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin*, *98*(2), 219-235.

Xia, Y., & Xu, F. (2022). Study on music emotion recognition based on the machine learning model clustering algorithm. *Mathematical Problems in Engineering*, *2022*(1), 9256586. https://doi.org/10.1155/2022/9256586.

Xiang, Z., Dong, X., Li, Y., Yu, F., Xu, X., & Wu, H. (2020). Bimodal emotion recognition model for Minnan songs. *Information*, *11*(3), 145. https://doi.org/10.3390/info11030145.

Xianyu, H., Li, X., Chen, W., Meng, F., Tian, J., Xu, M., & Cai, L. (2016). SVR based double-scale regression for dynamic emotion prediction in music. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 549-553). IEEE. Shanghai, China. https://doi.org/10.1109/icassp.2016.7471735.

Yang, J. (2021). A novel music emotion recognition model using neural network technology. *Frontiers in Psychology*, *12*, 760060. https://doi.org/10.3389/fpsyg.2021.760060.

Yang, X., Dong, Y., & Li, J. (2017). Review of data features-based music emotion recognition methods. *Multimedia Systems*, *24*(4), 365-389. https://doi.org/10.1007/s00530-017-0559-4.

Yang, Y., & Chen, H.H. (2011). Prediction of the distribution of perceived music emotions using discrete samples. *IEEE Transactions on Audio Speech and Language Processing*, *19*(7), 2184-2196.

Yang, Y., & Chen, H.H. (2012). Machine recognition of music emotion. *ACM Transactions on Intelligent Systems and Technology*, *3*(3), 1-30. https://doi.org/10.1145/2168752.2168754.

Yang, Y., Lin, Y., Su, Y., & Chen, H.H. (2008). A regression approach to music emotion recognition. *IEEE Transactions on Audio Speech and Language Processing*, *16*(2), 448-457.

Yang, Y.H., Liu, C.C., & Chen, H.H. (2006). Music emotion classification: a fuzzy approach. In *Proceedings of the 14th ACM International Conference on Multimedia* (pp. 81-84). ACM. https://doi.org/10.1145/1180639.1180665.

Yu, Z., Xu, X., Chen, X., & Yang, D. (2020). Learning a representation for cover song identification using convolutional neural network. In *IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 541-545). IEEE. Barcelona, Spain. https://doi.org/10.1109/icassp40776.2020.9053839.

Zentner, M., Grandjean, D., & Scherer, K.R. (2008). Emotions evoked by the sound of music: characterization, classification, and measurement. *Emotion*, *8*(4), 494-521. https://doi.org/10.1037/1528-3542.8.4.494.

Zhang, M., Zhu, Y., Ge, N., Zhu, Y., Feng, T., & Zhang, W. (2021). Attention-based joint feature extraction model for static music emotion classification. In *2021 14th International Symposium on Computational Intelligence and Design* (pp. 291-296). IEEE. Hangzhou, China. https://doi.org/10.1109/iscid52796.2021.00074.

Zuber, S., & Vidhya, K. (2022). Detection and analysis of emotion recognition from speech signals using decision tree and comparing with support vector machine. In *2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems* (pp. 1-5). IEEE. Chennai, India.