

Are We There Yet? A Brief Survey of Music Emotion Prediction Datasets, Models and Outstanding Challenges

Jaeyong Kang, Dorien Herremans *Senior Member, IEEE*

Abstract—Deep learning models for music have advanced drastically in recent years, but how good are machine learning models at capturing emotion, and what challenges are researchers facing? In this paper, we provide a comprehensive overview of the available music-emotion datasets and discuss evaluation standards as well as competitions in the field. We also offer a brief overview of various types of music emotion prediction models that have been built over the years, providing insights into the diverse approaches within the field. Through this examination, we highlight the challenges that persist in accurately capturing emotion in music, including issues related to dataset quality, annotation consistency, and model generalization. Additionally, we explore the impact of different modalities, such as audio, MIDI, and physiological signals, on the effectiveness of emotion prediction models. Through this examination, we identify persistent challenges in music emotion recognition (MER), including issues related to dataset quality, the ambiguity in emotion labels, and the difficulties of cross-dataset generalization. We argue that future advancements in MER require standardized benchmarks, larger and more diverse datasets, and improved model interpretability. Recognizing the dynamic nature of this field, we have complemented our findings with an accompanying GitHub repository¹. This repository contains a comprehensive list of music emotion datasets and recent predictive models.

Index Terms—Deep learning, artificial intelligence, music emotion recognition.

I. INTRODUCTION

Music has long been revered for its profound ability to evoke and convey emotions, transcending cultural, linguistic, and geographical barriers. Researchers from various fields have been captivated by the intricate interplay between music and human emotions for decades. Classic works such as Meyer's paper [1] and pioneering studies conducted by scholars such as Seashore [2] and Hevner [3] in the early 20th century laid the groundwork for understanding the emotional impact of music. However, understanding and quantifying this poses a significant challenge due to its multifaceted nature [4]. In this paper, we provide an overview of the current state-of-the-art along with challenges and directions for future work.

With the advent of technology and data-driven methodologies, new avenues have opened up for exploring the complex relationship between music and emotions, such as deep learning models. Despite these advancements, accurate predictions

J. Kang and D. Herremans are with the Information Systems Technology and Design Pillar at the Singapore University of Technology and Design (SUTD), Singapore 487372. E-mail: {jaeyong_kang, dorien_herremans}@sutd.edu.sg.

J. Kang is the corresponding author.

¹<https://github.com/AMAAI-Lab/awesome-MER/>

of emotions from music remain elusive. This is due to a number of reasons, including the subjective, personal nature of emotional perception, as well as bias and limitations in the current datasets, and challenges in benchmarking. We discuss these and other challenges at length in Section V.

Despite these challenges, the potential applications of Music Emotion Recognition (MER) systems are vast and varied. In the healthcare domain, we may find personalized music recommendation systems that guide a listener to different emotional states [5], or we may even see MER systems used to build large-scale datasets on which we can train generative music AI systems that can be controlled to generate music with specific emotions [6]–[8]. In the same line of thought, [9] developed a Brain-Computer Interface system that can provide musical feedback about the listener's current emotional state and subsequently influence this state. Additionally, MER systems may help facilitate emotional analysis during music composition, interactive experiences in media and entertainment, as well as inform market research [10]. The implications of understanding music and emotions extend across diverse domains. In general, following the idea of positive psychology [11], understanding music emotions can be used to enhance the user experience when designing various systems.

In this paper, we do not aim to provide a comprehensive overview of MER models, instead, we focus on discussing datasets, evaluation approaches, and identifying key challenges as well as future directions. We only briefly touch upon some of the more recent MER models. For a more comprehensive overview of MER machine learning models, the reader is referred to [10], [12]–[15].

Earlier surveys on Music Emotion Recognition (MER) include the review by Panda et al. (2020) [16], which focuses on the taxonomy of audio features and their relationship to emotion perception. More recently, a broad overview of deep learning techniques for MER was provided in [17], with extended discussion on emotion representations. However, these surveys primarily focus on either specific feature engineering strategies or model architectures, and they often lack comprehensive coverage of datasets, evaluation protocols, and the challenges of integrating multimodal sources or predicting induced emotion. In contrast, our goal is to provide a broader and more integrative perspective on the MER landscape. We examine not only the evolution of model architectures, but also the diversity of datasets, annotation strategies, emotion models, and evaluation practices. This comprehensive perspective allows us to highlight emerging trends, identify gaps in

TABLE I
OVERVIEW OF MUSIC EMOTION DATASETS (SORTED BY YEAR). NOTE: S = STATIC, D = DYNAMIC, B = BOTH; P = PERCEIVED, I = INDUCED.

Dataset	Year	# of instances	Length	Type	Categorical	Dimensional	S/D/B	P/I
MoodsMIREX [18]	2007	269	30s	MP3	5 labels	-	S	P
CAL500 [19]	2007	500	full	MP3	174 labels	-	S	P
Yang-Dim [20]	2008	195	25s	WAV	-	Russell	S	P
MoodSwings [21]	2008	240	15s	MP3	-	Russell	D	P
NTWICM [22]	2010	2,648	full	MP3	-	Russell	S	P
Soundtrack [23]	2011	470	15s-1m	MP3	6 labels	3 dimensions	S	P
MoodSwings Turk [24]	2011	240	15s	MP3	-	Russell	D	P
Last.fm subset of MSD [25]	2011	505,216	full	Metadata only	listener tags	-	S	P
DEAP [26]	2012	120	60s	YouTube id	-	Russell	S	I
Panda et al.'s dataset [27]	2013	903	30s	MP3, MIDI	21 labels	-	S	P
Solymani et al.'s dataset [28]	2013	1,000	45s	MP3	-	Russell	B	P
CAL500exp [29]	2014	3,223	3s-16s	MP3	67 labels	-	S	P
AMG1608 [30]	2015	1,608	30s	WAV	-	Russell	S	P
Emotify [31]	2016	400	60s	MP3	GEMS	-	S	I
Moodo [32]	2016	200	15s	WAV	-	Russell	S	P
Malheiros et al.'s dataset [33]	2016	200	30s	Audio, Lyrics	Quadrants	-	S	P
CH818 [34]	2017	818	30s	MP3	-	Russell	S	P
MoodyLyrics [35]	2017	2,595	full	Lyrics	4 labels	-	S	P
4Q-emotion [36]	2018	900	30s	MP3	Quadrants	-	S	P
DEAM [37]	2018	2,058	45s	MP3	-	Russell	B	P
PMEMo [38]	2018	794	full	MP3	-	Russell	B	I
RAVDESS [39]	2018	1,012	full	MP3, MP4	5 labels	-	S	P
DMDD [40]	2018	18,644	full	Audio, Lyrics	-	Russell	S	P
MTG-Jamendo [41]	2019	18,486	full	MP3	56 labels	-	S	P
VGMIDI [42]	2019	200	full	MIDI	-	Russell	D	P
Turkish Music Emotion [43]	2019	400	30s	MP3	4 labels	-	S	P
EMOPIA [44]	2021	1,087	30s-40s	Audio, MIDI	Quadrants	-	S	P
MER500 [45]	2020	494	10s	WAV	5 labels	-	S	P
Music4all [46]	2020	109,269	30s	WAV	-	3 dimensions	S	P
CCMED-WCMED [47]	2020	800	8-20s	WAV	-	Russell	S	P
MuSe [48]	2021	90,001	full	Audio	-	Russell (V-A-D)	S	P
HKU956 [49]	2022	956	full	MP3	-	Russell	S	I
MERP [50]	2022	54	full	WAV	-	Russell	B	P
MuVi [51]	2022	81	full	YouTube id	GEMS	Russell	B	P
YM2413-MDB [52]	2022	699	full	WAV, MIDI	19 labels	-	S	P
MusAV [53]	2022	2,092	full	WAV	-	Russell	S	P
EmoMV [54]	2023	5,986	30s	WAV	6 labels	-	S	P
Indonesian Song [55]	2023	476	full	WAV	3 labels	-	S	P
TROMPA-MER [56]	2023	1,161	30s	WAV	11 labels	-	S	P
Music-Mouv [57]	2023	188	full	Spotify id	GEMS	-	S	I
ENSA [58]	2023	60	full	MP3	-	Russell	D	P
EMMA [59]	2024	364	30s-60s	WAV	GEMS	-	S	I
SiTunes [60]	2024	300	full	WAV	-	Russell	S	I
MERGE [61]	2024	3,554	full	Audio, Lyrics	Quadrants	-	S	P
Popular Hooks [62]	2024	38,694	hooks	Video, Audio, Lyrics	Quadrants	-	S	P
Affolter and Rohrmeier's dataset [63]	2024	5,892	full	Spotify id	8 labels	-	S	P
XMIDI [64]	2025	108,023	full	MIDI	11 labels	-	S	P

current methodologies, and suggest promising directions for future research.

In the next section, we provide an extensive overview of the available emotion-annotated music datasets. This is followed by a discussion of the evaluation practices in the field (Section III). After that, we describe a selected list of recent models and approaches in Section IV. Finally, Section V dives into the remaining challenges and future direction for the field of music emotion recognition (MER), followed by a general conclusion.

II. DATASETS

Table I presents an extensive overview of emotion-annotated music datasets. These datasets vary in size, annotation granularity, and focus on either perceived or induced emotions. We

have attempted to provide an exhaustive overview of emotion datasets. This was achieved by using Google Scholar with search terms such as ‘music emotion dataset’, ‘affective music dataset’, as well as following references within these articles. Before delving deeper into the datasets, we first describe key dimensions that help structure our comparison. These include: (1) the emotion representation model (e.g., categorical vs. dimensional), (2) the annotation method (static or dynamic), (3) the type of emotion captured (perceived vs. induced), (4) the data modalities involved (e.g., audio, video, lyrics), and (5) the dataset’s size and diversity in genre or listener demographics. Below, we elaborate on each of these aspects and illustrate them with representative examples.

Emotion representations One of the earliest emotion representation models in music is Hevner’s affective ring [65],

developed in 1936. Based on extensive experimental studies, Hevner's model categorizes music emotions into eight fundamental categories: dignified, sad, dreamy, serene, graceful, happy, exciting, and vigorous. In current-day MER research, we see that Russell's Circumplex Model of Affect [66] is widely used. This model characterizes emotions along two dimensions: valence and arousal. Valence represents the degree of positive or negative emotion, while arousal reflects the intensity of emotion, ranging from passive to activated states. Russell's original model contained a third dimension: dominance [66]. This dimension is typically omitted as it can be hard to annotate, although some researchers have argued to re-include it [67]. It is worth noting, however, that the dominance dimension was more central to Russell's earlier PAD (Pleasure-Arousal-Dominance) framework, designed for environmental and contextual emotion analysis. Its omission from the Circumplex Model was a simplification intended to facilitate self-reported emotion studies.

Russell's model is a *dimensional* model, as it consists of continuous values along multiple dimensions (valence/arousal). Hevner's model, on the other hand, is *categorical* as it consists of discrete emotion labels. Thayer's two-dimensional model [68] focuses on energetic arousal and tense arousal as the primary dimensions of emotion. Thayer suggests that valence can be inferred from the combination of energetic and tense arousal levels. Other categorical models include the Geneva Emotional Music Scales (GEMS) [31]. This model was specifically designed for music-induced emotions and consists of 45 emotion tags grouped into nine categories, including amazement, solemnity, tenderness, nostalgia, calmness, power, joyful activation, tension, and sadness.

Recently, alternative emotion frameworks grounded in psychological theory have also been explored. For instance, Affolter et al. [63] introduced a dataset that maps listener-generated tags and playlist names to emotion labels derived from Plutchik's psychoevolutionary model of emotions. Their method employs natural language processing to associate each track with an 8-dimensional vector corresponding to Plutchik's basic emotions, offering a novel bridge between tag-based annotations and theoretical models of affect.

The datasets listed in Table I use a variety of emotion representation models, with Russell's model being the most popular dimensional representation. Some datasets do not use a specific emotion model, for instance, the MTG-Jamendo dataset [41] consists of freely assigned tags by the listeners, resulting in a diverse and comprehensive set of 56 tags, spanning from 'melancholic' to 'upbeat'. A subset of this dataset is used for the Emotion and Theme Recognition in Music' Task of MediaEval [69], which serves as a benchmark for evaluating MER systems. In Table I, the number of tags used to represent the emotions are listed in the column "Categorical".

Static versus dynamic Regardless of which emotion representation model is being used, we notice two fundamentally different approaches: static versus dynamic annotations. In a static setting, the listeners indicate the emotion for the entire song or fragment. In a dynamic annotation setting, the listener continuously indicates the emotion throughout the song or fragment. For instance, the MTG-Jamendo dataset

[41] offers categorical tags for each full-length song. The annotation is static, but multiple tags are allowed per song. The MoodSwings dataset [21], on the other hand, offers dynamic annotations of valence/arousal for every second of the musical fragments. Finally, some datasets offer both, for instance, MERP [50] provides both static GEM labels for the entire song, as well as dynamic valence/arousal ratings for every 1s of a song.

Induced versus perceived Emotion labels in the datasets can represent either perceived or induced emotions. Perceived emotions are emotions that listeners consciously recognize within the music itself. Induced emotions, on the other hand, are emotions that listeners experience as a result of listening to the music, which involve an actual emotional experience provoked by the stimulus. This emotional experience can be influenced by context, memories, and personal experiences, and may differ from the perceived emotion [70].

Most of the datasets in Table I use perceived emotion labels, which are easier to annotate. For induced emotion labels, researchers have used psycho-physiological measurements ranging from electromyogram (EMG), volume pulse (BVP), electrocardiograms (ECG), skin conductance, respiration rate, heart rate, to electroencephalograms (EEG). There have been many studies in psychology that use such biosensors to explore how music can influence our emotions [71]–[75]. Since these studies include medical data, the datasets are not often public. However, there are a few datasets with music and its induced emotions. Firstly the DEAP dataset [26] includes EEG, facial video recordings, as well as peripheral physiological signals that were recorded while they watched music videos. In addition, they collected perceived emotion ratings in terms of arousal, valence, like/dislike, dominance and familiarity. Second, the HKU956 dataset [49] records five kinds of physiological signals (i.e., heart rate, electrodermal activity, blood volume pulse, inter-beat interval, and skin temperature) of participants as they listen to music, along with reported emotions in the arousal and valence dimensions. Third, the Music-Mouv' dataset [57] investigates the impact of emotional context induced by music on gait initiation, focusing on anticipatory postural adjustments crucial for the elderly and individuals with Parkinson's disease. It includes subjective emotional responses, physiological data from wristbands, and biomechanical data from shoe insoles equipped with sensors collected during and after music listening. Finally, SiTunes [60] includes physiological signals (i.e., heart rate, activity intensity, activity step, and activity type) measured both before and after music listening, alongside environmental data (i.e., time of day, weather information, and location) recorded during users' daily lives. The MER models that can predict *induced* emotions have various medical applications, such as curating playlists to guide patients to different emotional states [5]. An interesting study by [76] concludes that music tends to induce the emotion that is perceived, enabling researchers to use perceived emotion datasets for developing emotion-inducing models.

Modalities Music comes in many formats, the most common one being 'audio'. Audio files can either be raw waveforms or compressed .mp3 files. However, we should not

neglect the MIDI format, a popular format still often used by music producers, composers and performers. Whereas most datasets contain audio files, only a handful focus on MIDI: 1) the VGMIDI dataset [42], which contains continuous valence/arousal ratings for MIDI files of piano arrangements for video game soundtracks; 2) the Panda et al.’s dataset [27], which provides a diverse collection of audio clips, lyrics, and aligned MIDI files, contains 28 emotion labels grouped into five emotion clusters derived from a cluster analysis of online tags by [18]; and 3) the EMOPIA dataset [44], which contains paired piano music audio with MIDI that has emotion annotations of the four high/low valence/arousal quadrants.

Many sources can elicit emotion. For instance, video, or lyrics may also affect our perceived or induced emotions. Some datasets offer alternative multimedia streams such as emotion-rated music videos from a variety of genres, including pop, rock, classical, and jazz as in the DEAP dataset [26]; or text of lyrics with the annotated music in the DMDD dataset [40]. The MuVi dataset [51] even offers isolated modality ratings for music videos. In this study, the raters were presented with either the music video, the music alone, or the muted video. The final dataset contains ratings for each of these modalities separately as well as together. This allows [51] to build a model on pure isolated modalities, which proved to be more accurate than a traditional model.

The RAVDESS dataset [39] is a multimodal dataset containing both speech and music with emotional expressions: calm, happy, sad, angry, fearful, surprise, and disgust for speech; and calm, happy, sad, angry, and fearful for music. It provides a rich combination of modalities, including Audio-only, Audio-Video, and Video-only formats.

Additionally, the EmoMV dataset [54] focuses on affective music-video correspondence learning by providing labeled pairs of music and video segments that either match or mismatch in terms of emotional content. This dataset enables models to learn the affective alignment between audio and visual modalities, making it particularly valuable for emotion-based matching and retrieval tasks.

Recently, the Popular Hooks dataset [62] was introduced as a multimodal dataset that contains 38,694 popular musical hooks (i.e., memorable sections of songs) with synchronized MIDI, music video, audio, and lyrics. It also offers detailed (predicted) labels for high-level musical attributes such as tonality, structure, genre, emotion, and region. Leveraging a pre-trained multimodal music emotion recognition framework, the dataset provides predicted emotion labels, which were evaluated through a user study.

Lastly, datasets like MERP [50] and EMOPIA [44] provide metadata on song features and listener demographics, which can be useful for detailed analysis and personalized emotion prediction.

Dataset size Compared to affective datasets available in other domains (e.g., the Sentiment140 dataset [77] in NLP, which contains 1.6 million tweets annotated as positive or negative), the size of the available datasets is still very limited, with the largest dataset containing 109,269 instances. In addition to the number of instances in datasets, we also notice a difference in the length of the instances. A number of

datasets (e.g. MERP [50], VGMIDI [42], and HKU956 [49]) offer ratings for full-length songs. In datasets with dynamic ratings throughout the song, this may provide a means for researchers to analyse how our emotions evolve throughout a song. Other datasets focus on short fragments, often ranging from 30 seconds to 1-minute fragments (e.g. DEAM [37], EMOPIA [44], and EMMA [59]), but even as short as 10s as is the case for the MER500 dataset which consists of Indian Hindi film music [45].

Variety of music The genres covered in these datasets vary widely, reflecting the diverse nature of music. For instance, the DEAP dataset [26] includes classical as well as jazz pieces, which are known for their rich emotional and structural complexities. The VGMIDI dataset on the other hand, [42] focuses exclusively on video game soundtracks, a genre that often aims to evoke specific emotions to enhance the gaming experience. Rock and electronic music can primarily be found in the DEAM dataset [37], both genres that are prevalent in contemporary music culture. This genre diversity allows researchers to select datasets that align with their specific research goals and to explore how different genres affect emotional perception and annotation. Additionally, studying a variety of genres can help in developing more robust and generalized MER models that can perform well across different types of music.

Annotation Process Annotation processes and label distributions vary significantly across datasets, impacting the effectiveness and reliability of machine learning models in MER. Some datasets, such as MTG-Jamendo [41], rely on freely assigned tags by multiple annotators, allowing a broad range of emotional descriptors. In contrast, DEAP [26] and HKU956 [49] use a predetermined list of emotions to ensure consistency across annotations.

A major distinction between annotation strategies relates to if they use absolute or relative annotation methods, this is particularly important to note for dimensional models of emotion. In absolute annotation, annotators assign a direct score or category to each sample independently. This approach is common in datasets such as DEAP and HKU956, where participants provide valence and arousal ratings individually for each stimulus. Recently, several datasets propose *relative* annotation strategies to simplify the annotation process and improve consistency across annotators. Instead of rating individual tracks absolutely, annotators are asked to compare pairs of samples and judge their relative emotional positioning. Examples include the MusAV dataset [53], which gathers comparative annotations of arousal and valence for pairs of music tracks, and Emo-Soundscapes [78], where participants rank soundscape recordings based on perceived emotion. Earlier work by Yang et al. [79] introduced ranking-based emotion recognition methods for music organization and retrieval, showing that relative judgments can lower cognitive load and lead to more reliable ground truth data. Similarly, in the EMusic dataset [80], a ranking-based annotation approach was used for experimental music, demonstrating improvements over traditional absolute labeling methods.

In addition, the CCMED-WCMED dataset [47] compared Western and Chinese classical music based on emotion dimen-

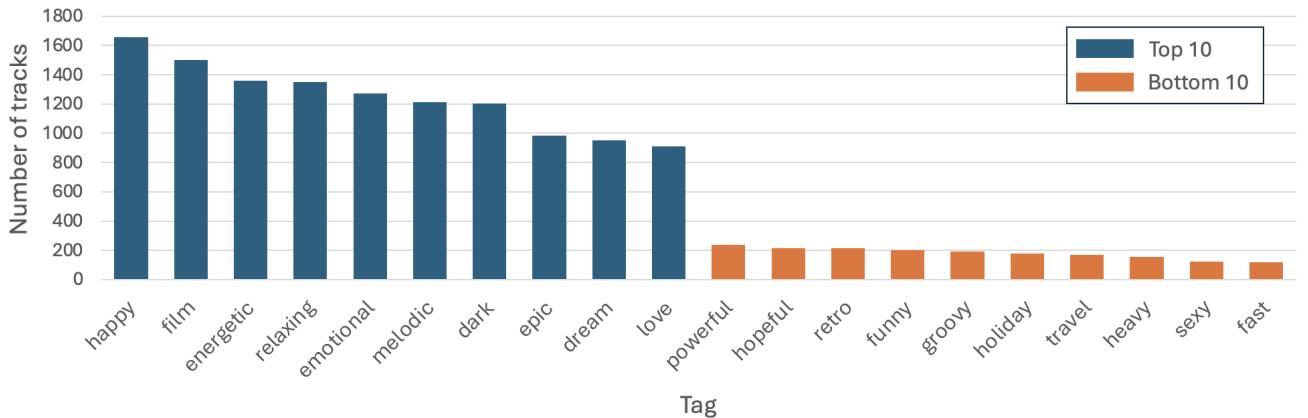


Fig. 1. Top 10 most and least frequent tags in the MTG-Jamendo dataset.

sions collected through relative annotations, highlighting the importance of cultural factors in emotion perception. These relative approaches can mitigate some of the variability and subjectivity inherent in absolute annotations, particularly in continuous emotion models like valence-arousal space.

The number of annotators also varies: some datasets involve large-scale crowdsourcing efforts, providing a wide range of perspectives (e.g., MusAV, Emo-Soundscapes), while others use smaller, more controlled groups to maintain annotation consistency (e.g., DEAP, HKU956). Additionally, the logic behind selecting music segments differs, with some datasets focusing on full-length songs and others on shorter clips to better capture transient emotional responses.

Overall, the choice between absolute and relative annotation methods, the number of annotators, and the segment selection strategies are all critical factors influencing dataset reliability and the resulting model performance in MER research.

Label distribution and splits The distribution of labels across datasets can be imbalanced, affecting the performance of machine learning models. For instance, the MTG-Jamendo dataset [41] contains a diverse set of 56 tags, freely assigned by users, but their frequency is highly imbalanced. This skewed distribution, with some tags like “happy” and “film” appearing far more often than others like “sexy” or “fast”, is visualized in Figure 2, which shows the top 10 most and least frequent tags. In contrast, the DEAP dataset [26] provides a more balanced set of labels across arousal and valence dimensions. Finally, some datasets, such as DEAM [37] and MTG-Jamendo [41], propose fixed splits for training and testing. This facilitates the replication of experiments and ensures consistent evaluation metrics which are easily used as a benchmark.

Recommendations and Future Directions While the landscape of emotion-annotated music datasets is expanding, several challenges and opportunities remain. For general-purpose MER models, datasets like MTG-Jamendo [41] and MERP [50] offer breadth in genre and label diversity, while EMOPIA [44] and VGMIDI [42] are more suitable for symbolic-domain or piano-focused emotion modeling. For multimodal applications, Popular Hooks [62] and EmoMV [54] are relevant. However, there is a notable lack of large-scale, publicly available datasets focusing on induced emotions with

multimodal or biosensor data, limiting medical and affective computing applications. The community would benefit from more datasets capturing diverse demographics, cross-cultural emotional interpretations, and consistent emotion annotation schemas across modalities. We encourage future dataset creators to include dynamic annotations, listener metadata, and align modalities when possible.

In sum, there is a variety of emotion-annotated datasets available as shown in Table I. They differ in terms of emotion representation models, annotations labels, as well as music format and perceived versus induced emotion labels. Understanding these variations is crucial for developing robust and accurate MER models that can generalize across different datasets and musical genres. The challenges and future research directions related to datasets are discussed in detail in Section V.

III. EVALUATION PROTOCOLS

Evaluation metrics play a crucial role in assessing the performance of MER systems. Depending on which type of emotion ratings are being predicted (dimensional versus categorical), the evaluation metrics change as we are dealing with a regression or a classification task respectively. Commonly used evaluation metrics for categorical MER systems include accuracy, precision, area under the ROC curve (AUC), and confusion matrices. In the case of regression MER models, metrics such as mean squared error (MSE), R^2 , and Pearson correlation coefficient [81] are used. These metrics provide insights into the effectiveness of MER models when it comes to emotions represented by dimensional models.

The evaluation approach also depends on whether a dataset provides static or dynamic annotations. Static annotations assign a single label (or vector) to an entire track or segment, making evaluation straightforward with standard classification or regression metrics (e.g., accuracy, F1-score, ROC-AUC, or R^2) computed at the song level [20], [36], [41]. In contrast, dynamic annotations provide time-continuous emotion ratings—often sampled at a rate of 1 Hz—reflecting how emotions evolve throughout a track [26], [37], [38]. In such cases, evaluation typically involves comparing predicted and ground-truth time series using frame-level metrics such as

Pearson Correlation Coefficient (PCC), Mean Squared Error (MSE), or Concordance Correlation Coefficient (CCC) [37], [38], [82]. Some studies apply smoothing or window-based aggregation (e.g., moving average or median filters) to reduce annotation noise and capture broader emotional trends [37], [82]. For example, the DEAM dataset [37] provides both static and dynamic valence-arousal annotations, with dynamic evaluation conducted on a per-frame basis across time.

When evaluating MER models, it is also important to distinguish between perceived and induced (invoked) emotion. Evaluating perceived emotion predictions typically involves comparing predicted labels with human-annotated ground truth, which reflects how listeners interpret the emotional content of the music. In contrast, evaluating induced emotion is more complex, as it involves the actual emotional response elicited in the listener. This often requires physiological measurements (e.g., heart rate, EEG) or self-reports in a controlled experimental setup. For this reason, evaluation of induced emotion prediction models may rely on additional modalities (e.g., biosensors), and often focuses on subjective measures such as emotion regulation effectiveness or user satisfaction. These differences highlight the need for tailored evaluation protocols depending on the target emotion type.

Evaluating MER systems goes beyond just defining a common metric. Due to the inherent differences between datasets, a comparison across datasets is often not possible. For an in-depth discussion on this, the reader is referred to Section V. Within one dataset, however, it is possible to establish benchmarks and compare the performance of different models, provided that the train/test split is shared. There are some initiatives to facilitate the comparison between models, such as competitions, as well as individual papers that offer clear data splits and metrics [82]–[86].

Competitions and challenges provide valuable platforms for evaluating and comparing the performance of MER systems. These initiatives often involve standardized datasets, evaluation protocols, and metrics, enabling researchers to benchmark their algorithms against state-of-the-art methods. Existing benchmarking initiatives and competitions in MER include the ‘Audio K-POP Mood Classification’ task in MIREX (Music Information Retrieval Evaluation eXchange) (last organized in 2019)², the ‘Emotion in Music’ task in MediaEval (last organized in 2015)³, and the ‘Emotion and Theme Recognition in Music Using Jamendo’ in MediaEval (last organized in 2021)⁴.

Given the variability and potential noisiness of emotion annotations in music datasets [50], evaluation protocols should be designed to assess not only performance against possibly imperfect ground truth but also the perceptual validity of model outputs. One way to address this is by incorporating listening tests, where human subjects rate the emotions conveyed by model-predicted tracks. This does not directly mitigate the noise in the training data, but rather provides

²https://www.music-ir.org/mirex/wiki/2019:Audio_K-POP_Mood_Classification/

³<http://www.multimediaeval.org/mediaeval2015/emotioninmusic2015/>

⁴<https://multimediaeval.github.io/2021-Emotion-and-Theme-Recognition-in-Music-Task/>

an additional, human-centered validation of the model’s effectiveness. Another complementary evaluation strategy is extrinsic evaluation, where emotion recognition is applied in downstream tasks. For instance, predicted emotions could be used to generate emotionally coherent playlists, and users could be asked whether the playlists evoke the intended emotional responses. Such evaluations help assess whether MER systems produce outputs that are meaningful and useful in real-world applications, beyond what can be inferred from standard metrics alone.

IV. MODELS AND APPROACHES

In this section, we briefly touch upon some of the more recent MER models, highlighting the state-of-the-art approaches over the last few years. This is not an exhaustive overview, and for a more comprehensive review, readers are referred to other survey papers [10], [12]–[15]. The aim of this section is to point out the *current state-of-the-art* and various approaches over the last five years. The models were selected through Google Scholar searches using the search terms ‘music emotion prediction model’, ‘affective music prediction model’, and ‘music emotion recognition model’ starting from the year 2020, as well as by following links in the articles.

Table II presents an overview of selected MER models released since 2020, summarizing their modalities, approaches, emotion models, and datasets. This table highlights the diversity in methodologies and datasets used in recent research.

Some of the earliest attempts at music emotion prediction involved rule-based approaches and hierarchical frameworks. For instance, Feng et al. [125] used Computational Media Aesthetics (CMA) to analyze tempo and articulation, mapping them into four mood categories: happiness, anger, sadness, and fear. They achieved a total precision of 67% and a total recall of 66%. Lu et al. [126] developed a hierarchical framework for automatically extracting music emotion from acoustic data. They employed music intensity to represent the energy dimension of Thayer’s model while using timbre and rhythm to capture the stress dimension. They achieved an average accuracy of mood detection of up to 86.3%. These early models laid the groundwork for subsequent advancements in music emotion recognition, paving the way for the adoption of more sophisticated techniques, including deep learning approaches.

In recent years, various deep-learning MER models have been developed, employing techniques such as Convolutional Neural Networks (CNNs) [55], [83], [88], [90], [92]–[94], [96]–[98], [103], [109], [112], [116], [119], [121], [122], Recursive Neural Networks (RNNs) such as Gated Recurrent Units (GRUs) [112] and Long-Short Term Memory networks (LSTMs) [51], [55], [88]–[91], [96], [97], [99], [109], [118], and more recently, Transformer architectures [55], [100], [114], [117].

Additionally, generative models such as Generative Adversarial Networks (GANs) have been applied. For instance, Huang et al. [102] developed a GAN-based model for emotion recognition using audio inputs under IoT environments.

On the MTG-Jamendo dataset [41], several models have demonstrated noteworthy performance. For example, Mayerl

TABLE II
OVERVIEW OF SELECTED MUSIC EMOTION RECOGNITION MODELS SINCE 2020. NOTE: V = VALENCE, A = AROUSAL, Q = QUADRANT.

Ref.	Year	Modalities	Approach	Emotion Model	Dataset	Performance
[87]	2020	Audio, Lyrics	Machine Learning (e.g., SVM, NB)	Russell	PMEmo	Accuracy: 63%
[88]	2020	Audio	CNN, LSTM	Russell	Solymani et al.'s dataset	RMSE: 0.219 V, 0.212 A
[89]	2020	Audio	LSTM	happy, sad, neutral, fear	Self-built	Accuracy: 89.3%
[90]	2020	Audio, Lyrics	CNN-LSTM	angry, happy, relaxed, sad	Last.fm	Accuracy: 78%
[91]	2020	Audio	Attentive LSTM	Russell	Solymani et al.'s dataset	R2: 0.53 V, 0.75 A
[92]	2020	Audio	Source Separation, CNN	Russell	PMEmo	R2: 0.4814 V, 0.6004 A
[93]	2020	Audio	Cochleogram, CNN	Russell	Solymani et al.'s dataset	R2: 0.41 V, 0.63 A
[94]	2020	Audio	CNN, Local Attention	Quadrants	Soundtrack, Bi-Modal	Accuracy (Soundtrack): 67.71% F1 (Bi-Modal): 77.82%
[83]	2020	Audio	Attention-based Neural Networks	56 mood/theme tags	MTG-Jamendo	PR-AUC: 0.118, ROC-AUC: 0.735
[82]	2020	Audio	Triplet Neural Networks	Russell	Solymani et al.'s dataset, DEAM	R2 (Solymani.): 0.378 V, 0.638 A R2 (DEAM): 0.361 V, 0.672 A
[95]	2020	Audio	CNN Ensemble, Focal Loss, Receptive Field Tuning	56 mood/theme tags	MTG-Jamendo	PR-AUC: 0.161, ROC-AUC: 0.781
[96]	2021	Audio	CNN-LSTM	Russell	DEAM	Accuracy: 64.9%
[97]	2021	Audio	CNN, LSTM+DNN	3 classes (high arousal + pos./neg. valence)	Self-built	Accuracy: 99.19%
[84]	2021	Audio	Clustering-based Ensembles	56 mood/theme tags	MTG-Jamendo	PR-AUC: 0.109, ROC-AUC: 0.705
[85]	2021	Audio	Semi-Supervised, Noisy Student Training	56 mood/theme tags	MTG-Jamendo	PR-AUC: 0.136, ROC-AUC: 0.769
[86]	2021	Audio	Frequency Dependent Convolutions	56 mood/theme tags	MTG-Jamendo	PR-AUC: 0.151, ROC-AUC: 0.775
[98]	2021	Audio	CNN, Co-teaching Training Strategy	56 mood/theme tags	MTG-Jamendo	PR-AUC: 0.144, ROC-AUC: 0.760
[99]	2021	Audio	LSTM, Pretrained Models	Russell	Self-built	R2: 0.46 V, 0.73 A
[100]	2021	Lyrics	Transformers	Russell	MoodyLyrics, MER	Accuracy (MoodyLyrics): 94.78% Accuracy (MER): 94.44% V, 88.89% A, 88.89% Q
[101]	2021	Audio	Source-separation-based Explainer	Russell	DEAM, Midlevel, PMEMo	R2 (DEAM): 0.48 V, 0.50 A R2 (PMEmo): 0.50 V, 0.65 A
[102]	2021	Audio	Generative Adversarial Network	happy, sad	Self-built	Accuracy: 87%
[103]	2021	Audio, Video	CNN	exciting, fear, neutral, relaxation, sad, tension	Self-built	Accuracy: 88.56%
[104]	2021	Audio	Linear Regressors	Russell	Self-built	R2: 0.78 V, 0.85 A
[105]	2022	Audio	Clustering, Machine Learning (e.g., SVM)	Russell	Solymani et al.'s dataset	Accuracy: 84.9%
[106]	2022	MIDI	Multi-Task Learning	Quadrants	EMOPIA, VG-MIDI	Accuracy (EMOPIA): 67.58% Accuracy (VG-MIDI): 55.85%
[107]	2022	Audio	Feature Selection, SVR, RF	Russell	DEAM	R2: 0.587 V, 0.645 A
[51]	2022	Audio, Video	LSTM	Russell	Self-built	RMSE: 0.1331 V, 0.0973 A
[108]	2022	Lyrics	Deep Learning, BERT	Quadrants	MIR Lyrics	F1: 88.9 %
[109]	2022	Audio	CNN, LSTM	Russell	Emotion PMEMo, AllMusic	Accuracy (PMEmo): 79.01% V, 83.62% A Accuracy (AllMusic): 67.11% V, 86.56% A
[110]	2022	Audio	SVM, Random Forest, MLP	Russell	DEAM	RMSE: 0.23 V, 0.24 A
[111]	2023	Audio, Lyrics	Multi-Modality	Russell	DMDD	R2: 0.235 V, 0.196 A
[112]	2023	Audio	Inception-GRU Residual	Quadrants	Soundtrack	Accuracy: 84%
[113]	2023	Lyrics	State Space Models	anger, disgust, fear, joy, sadness, surprise	LyricsEmotions	Pearson correlation: 0.345 (angry), 0.268 (disgust), 0.350 (fear), 0.503 (joy), 0.350 (sad), 0.089 (surprise)
[114]	2023	Lyrics	Fine-tuned XLMRoBERTa	Russell	Self-built	Accuracy: 80.88% V, 81.51% A, 62.58% Q
[55]	2023	Audio, Lyrics	CNN-LSTM, XLNet Transformers	positive, neutral, negative	Self-built	Accuracy: 80.56%
[115]	2023	Audio	Attention Mechanism	Russell	DEAM, PMEMo	RMSE (DEAM): 0.112 V, 0.109 A RMSE (PMEmo): 0.144 V, 0.135 A
[116]	2023	Audio	CNN	happiness, fear, sadness, peacefulness	Musical Excerpts	Accuracy: 92%
[117]	2023	Audio, Lyrics	Transformers	Quadrants	MoodyLyrics	Accuracy: 77.94%
[118]	2024	Audio	RNN, BRNN, LSTM	Quadrants	4Q audio, MTG-Jamendo	Accuracy (4Q.): 65.97% Accuracy (MTG.): 53.97%
[119]	2024	Audio, Lyrics	VGGish, ALBERT	happy, sad, calm, healing	DEAM, FMA	Accuracy (DEAM): 49.68% Accuracy (FMA): 49.54%

TABLE II
(CONTINUED)

Ref.	Year	Modalities	Approach	Emotion Model	Dataset	Performance
[120]	2024	Audio, Lyrics	Stacked Ensemble Models	angry, happy, relax, sad	Self-built	Accuracy: 96.25%
[121]	2024	Audio	Multi-scale Parallel Convolution	sleepy, calm, sad, pleased, relaxed, nervous, annoying, excited	PMEmo, Soundtrack, RAVDESS	Accuracy: 98.58%
[122]	2024	Audio	CNN with Differential Evolution	happy, sad, angry, calm	self-built, DEAM	Accuracy: 85.29%
[123]	2024	Audio	LLM Embeddings, Non-parametric Clustering	Categorial	MTG-Jamendo, CAL500, Emotify	F1 (MTG-Jamendo): 2.02% F1 (CAL500): 22.9% F1 (Emotify): 40.0%
[124]	2025	Audio	MERT Embeddings, Multitask Learning, Transformers, Knowledge Distillation	Categorial, Russell	MTG-Jamendo, PMEMo, DEAM, Solymani et al.'s dataset	(MTG-Jamendo) PR-AUC: 0.1543 ROC-AUC: 0.7810 (Solymani.) R2: 0.6512 V, 0.7616 A (PMEmo) R2: 0.5473 V, 0.7940 A (DEAM) R2: 0.5184 V, 0.6228 A

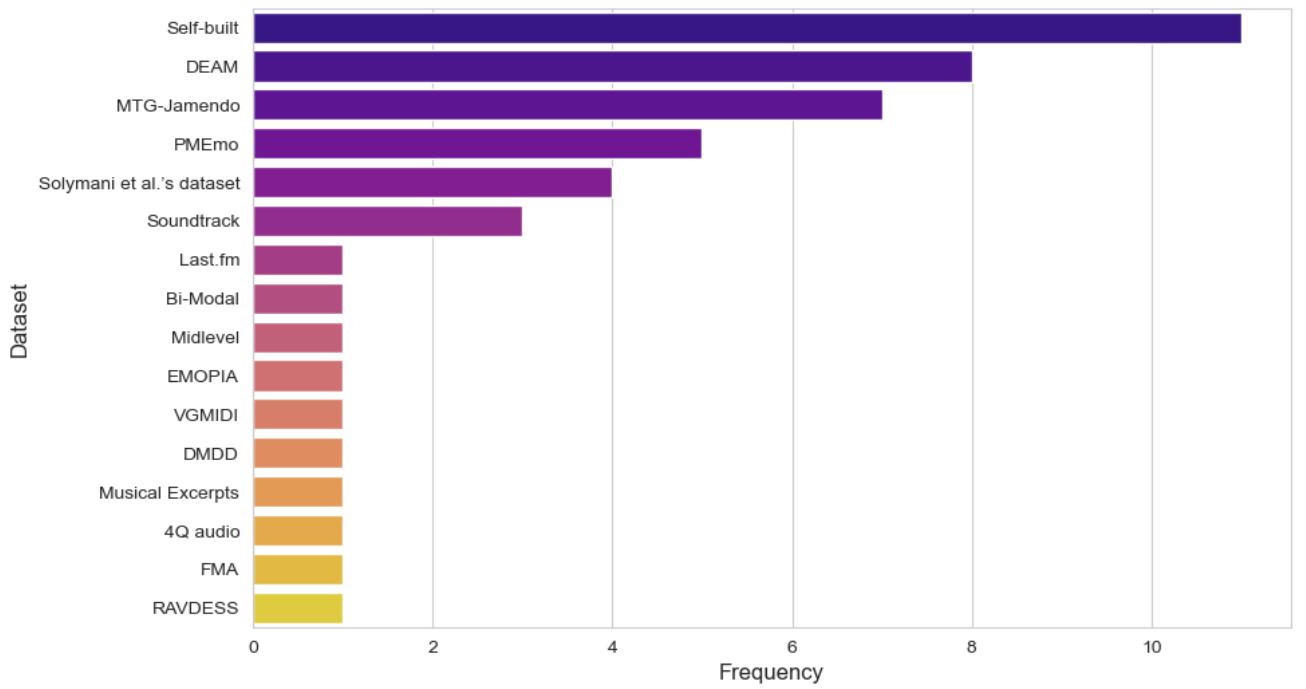


Fig. 2. Number of Music Emotion Recognition Models since 2020 that use the listed datasets.

et al. [84] employed a CNN-based (ResNet18) model using clustering-based ensembles, achieving a PR-AUC-macro score of 0.1087 and a ROC-AUC-macro score of 0.7047. Tan et al. [85], using a semi-supervised learning method and a CRNN architecture, reported higher scores of PR-AUC-macro 0.1357 and ROC-AUC-macro 0.7687. Pham et al. [98] achieved similar results with an EfficientNet-based CNN model, utilizing a co-teaching strategy to manage noisy data, resulting in PR-AUC-macro 0.1435 and ROC-AUC-macro 0.7599. Bour et al. [86] leveraged frequency-dependent convolutions in a CNN model, achieving PR-AUC-macro 0.1509 and ROC-AUC-macro 0.7748. Kang and Herremans [124] proposed a unified multitask learning framework that leverages both categorical and dimensional emotion labels by combining musical features

(e.g., key and chords) with MERT representations. Through knowledge distillation across multiple datasets, this approach enhances cross-dataset generalization and achieves improved performance on MTG-Jamendo with a PR-AUC of 0.1543 and ROC-AUC of 0.7810. Knox et al. [95] proposed an ensemble of CNN-based models trained on Mel spectrograms, exploring the impact of different loss functions and resampling strategies for multi-label music tagging. They showed that using focal loss effectively addressed class imbalance, and adjusting the CNN's receptive field further improved performance. Their model achieved the highest reported scores in the MediaEval "Emotion and Theme Recognition in Music" task, with a PR-AUC-macro of 0.1610 and ROC-AUC-macro of 0.7810.

Notably, hybrid architectures combining multiple deep

learning techniques, such as CNN-LSTM models, have shown promising results. For instance, Chen and Li [90] employed a CNN-LSTM model for emotion recognition using the Last.fm subset of the Million Song Dataset—a large-scale collection of music metadata with listener-generated tags. Their model outperformed standalone CNN and LSTM baselines, achieving an accuracy of 68.1% for audio classification and 74.2% for lyric classification. In contrast to classification approaches, He et al. [88] proposed a regression-based model using multi-view CNNs and bidirectional LSTMs directly on raw audio. Without relying on handcrafted features, their model achieved strong performance with RMSE scores of 0.219 for valence and 0.212 for arousal on the Solymani et al.’s dataset [28]. Transformer-based architectures have also gained traction in recent years, with notable examples by Suresh et al. [117] and Agrawal et al. [100], who explored their effectiveness in MER. Suresh et al., for example, applied a Transformer model to the MoodyLyrics dataset [35], which contains 2,595 tracks labeled with four mood categories derived from lyrics. Their model achieved the highest reported accuracy of 77.94%, outperforming CNN and Bi-GRU baselines.

Other models have employed more traditional machine learning techniques alongside deep learning approaches. Medina et al. [110] focused on classifying emotions from audio using SVM, Random Forest, and Multi-Layer Perceptron (MLP) models, achieving an F-score of 0.73 and 0.69 for predicting valence and arousal values, respectively, on the DEAM dataset [37]. Sharma et al. [87] combined machine learning algorithms such as Support Vector Machines (SVM) and Naive Bayes (NB) to predict emotions using both audio and lyric features. They achieved an accuracy of 63% on the PMEMo dataset [38]. Similarly, Griffiths et al. [104] developed a multi-genre MER model using linear regressors. They achieved a R^2 score of 0.776 and 0.85 for predicting valence and arousal values, respectively, on the self-built dataset. Meanwhile, Xia et al. [105] utilized clustering algorithms alongside machine learning techniques like SVM and K-Nearest Neighbors (KNN) for emotion recognition, and their hybrid model which combined all the machine learning techniques achieved an accuracy of 85% on the Solymani et al.’s dataset [28].

A distinction between the various models can be made based on the input modality. Some models are exclusively based on MIDI such as the multi-task architecture proposed by [106], and as such use a token-based representation. Given the limited size of MIDI datasets, the accuracy of such models is limited, e.g., the model by [106] reaches 67.56% accuracy when predicting between four emotion classes. Most of the existing MER models are based on audio, and hence they take as input raw audio. This is often converted into Mel-spectrograms which are then processed through convolutional neural networks (CNNs) [94], or the audio could be directly processed through a WaveNet architecture, which is a type of temporal CNN [127].

In recent years, audio embeddings pretrained on large-scale datasets have become increasingly common, enabling transfer learning for Music Emotion Recognition (MER). Beyond traditional embeddings, recent research has explored both supervised and unsupervised pretraining strategies for

audio representation learning, where MER is treated as a downstream task. For instance, MERT (Music undERstanding model with self-supervised Training) [128] is specifically tailored to music, addressing challenges like pitch and tonality through a novel training scheme. MERT incorporates pseudo labels from two teacher models—an acoustic teacher based on Residual Vector Quantization Variational AutoEncoder (RVQ-VAE) and a musical teacher based on the Constant-Q Transform (CQT)—within a masked language modeling framework. When used as a feature extractor for MER, MERT-95M achieves a PR-AUC of 0.1340 and ROC-AUC of 0.7640 on the MTG-Jamendo dataset [41], while MERT-330M reaches a PR-AUC of 0.1400 and ROC-AUC of 0.7650. Building upon MERT embeddings, Kang and Herremans [124] introduce a unified multitask learning framework that leverages both categorical and dimensional emotion labels. Their model achieved a PR-AUC of 0.1543 and ROC-AUC of 0.7810 on the MTG-Jamendo dataset [41]. For dimensional emotion regression, it attained R^2 scores of 0.5473 for valence and 0.7940 for arousal on PMEMo [38], 0.5184 for valence and 0.6228 for arousal on DEAM [37], and 0.6512 for valence and 0.7616 for arousal on the dataset by Solymani et al. [28].

McCallum et al. [129] present a comparative analysis of audio representation learning strategies, showing that supervised pretraining on expert-annotated music datasets leads to state-of-the-art performance in emotion tagging. Their models achieve 78.6 ROC-AUC and 16.1 PR-AUC on the MTG-Jamendo Mood/Theme dataset—outperforming previously reported benchmarks. Alonso-Jiménez et al. [130] propose MAEST, a transformer-based architecture trained with patchout and pre-initialized with ImageNet or AudioSet weights. Although its performance is slightly lower (78.1 ROC-AUC and 15.4 PR-AUC), MAEST demonstrates the potential of efficient, convolution-free architectures in music emotion tagging, particularly when speed and scalability are important.

Alternatives to these approaches include directly extracting spectral features (e.g., MFCCs, spectral centroids) with libraries such as OpenSmile, which has a configuration file specifically for the emotion recognition task [131]. [82] uses this approach and achieves R^2 scores of 0.378 and 0.638 for predicting dynamic valence and arousal values, respectively, on the Solymani et al.’s dataset [28]. Recognizing the importance of induced emotions, recent research has focused on models that leverage physiological data to predict music-induced emotions. For instance, [49] constructed the HKU956 dataset [49] with aligned peripheral physiological signals (i.e., heart rate, skin conductance, blood volume pulse, skin temperature) and self-reported emotion from 30 participants. The study revealed that physiological features significantly contribute to valence classification and that multimodal classifiers outperform single-modality ones. [101] uses PMEMo [38] as induced emotion recognition dataset. This study merges audioLIME, a source-separation-based explainable model, with mid-level perceptual features to form an intuitive connection between input audio and emotion predictions, providing insights into model predictions. [115] also uses PMEMo as induced emotion recognition dataset and introduces a novel

method named Modularized Composite Attention Network (MCAN). This method enhances feature extraction and employs attention mechanisms to improve the stability and accuracy of emotion prediction models.

Finally, musically meaningful features may be included, such as Rhythmic features (e.g., tempo, beat histogram), or note features (e.g., pitch), as implemented by Shi et al. [132] and Panda et al. [36], respectively. Shi et al. achieved a precision of 92.8% for 4 emotion categories (calm, sad, pleasant, and excited), while Panda et al. achieved an F1-score of 76.0% for 4 emotion categories (Quadrants).

Other input modalities may include text (lyrics), or video. In the case of the former, some models extract the sentiment from the lyrics using Natural Language Processing (NLP) tools [108], [111], or they use the entire lyrics with an embedding model [113], [117]. These features are then combined with audio-based features or analyzed independently to predict the emotional content of music. Including the lyrics does not always improve the sentiment prediction, particularly in predicting arousal [111]; however, [120] did manage to increase the performance of a MER model by using embeddings, such as word2vec [133] or stacked ensemble models that integrate both audio and lyrics. Finally, models that include video modalities typically use pretrained networks to capture image and video features, and thus improve model performance. For instance, [54] uses ResNet-50 [134] and FlowNetS [135], respectively.

Identifying the current state-of-the-art Music Emotion Recognition (MER) model is challenging due to factors such as differences in datasets and performance metrics, as highlighted in the previous section. A commonly used benchmark is the Emotion and Theme Recognition in Music competition based on the MTG-Jamendo dataset, where the top-performing model by Knox et al. [95] achieved a PR-AUC-macro of 0.161 and a ROC-AUC-macro of 0.781 using an ensemble of CNN-based models trained with focal loss and receptive field tuning. While benchmark performances of MER models have steadily improved (e.g., some achieving ROC-AUC scores over 0.78 on datasets like MTG-Jamendo), their translation to real-world applications remains limited.

One area where we do see emotion models integrated is emotion-conditioned music generation, where preliminary MER models are incorporated to guide the emotional content of generated music (e.g., Makris et al. [6]). Furthermore, emotion detection systems have been piloted in adaptive gaming environments or personalized music therapy applications [5], but large-scale real-world deployment of deep MER models remains rare due to variability across listeners, subjective annotation challenges, and differences between training and inference conditions.

Figure 2 illustrates the most frequently used datasets by the listed Music Emotion Recognition (MER) models since 2020, as summarized in Table II. This count includes studies where MER is the primary task as well as those where it serves as a downstream application, such as in music representation learning [128]. Self-built datasets, created by authors for specific research purposes, remain the most commonly used, appearing in 11 instances. These datasets often contain

licensed music, restricting their accessibility to the broader research community. Other frequently utilized datasets include DEAM (8 occurrences) and MTG-Jamendo (7 occurrences), highlighting their importance in MER studies. Some datasets, such as PMEMo [38] and Solymani et al.’s dataset [28], appear less frequently but still contribute valuable insights. For instance, PMEMo [38] includes dynamic emotion labels and physiological signals (EDA), enabling multimodal affective analysis, while Solymani et al.’s dataset [28] offers continuous valence-arousal annotations and standard deviations, supporting studies on annotation reliability and temporal emotion dynamics.

We have only provided a glimpse into the existing MER models in this section. From the performance of the various models, we see, however, that much improvement can still be made. We discuss some of the remaining challenges in the next section.

V. CHALLENGES AND FUTURE DIRECTIONS

Whereas the first publications on music and emotion surfaced in the 1930s [65], current MER models still struggle to match human performance in emotion recognition. The field of MER still faces several challenges and various opportunities for future exploration remain, ranging from overcoming data limitations to the integration of emerging technologies. Understanding and addressing these challenges is crucial for advancing the field.

Dataset limitations One of the primary challenges in MER is the scarcity of large, diverse, copyright-cleared, emotion-annotated datasets. Limited datasets hinder the development and evaluation of robust MER models, leading to potential biases and generalization issues, and also prevent the establishment of reliable performance benchmarks across studies.

We have recently seen advances in this area with the release of larger datasets such as MTG-Jamendo [41], Music4all [46], MuSe [48], which contain 18k, 109k, and 90k instances respectively. Whereas MTG-Jamendo and MuSe are available under a Creative Commons licence, however, Music4all contains copyrighted tracks. In addition, when exploring datasets, we notice that they are often skewed towards one particular genre. For instance, the DEAM dataset [37] consists mostly of rock and electronic music genres, while the VGMIDI dataset is focused solely on video game soundtracks.

To deal with the current dataset size limitations, techniques such as unsupervised pretrained may be helpful. With this technique, latent representations are first learned using unlabeled datasets. This evolution has led to some available large-scale audio encoders such as the Variational Auto Encoder used in [136], and the AST Audio Spectrogram Transformer presented in [137], as well as various recently developed neural audio encoders (e.g. Descript Audio Codec [138]). These novel pretrained representations may help deal with the limiting size of emotion-annotated datasets.

Subjective labels The subjective and variable nature of emotion perception also poses a significant challenge. Emotions are inherently complex and subjective, varying across individuals, cultures, and contexts [50]. For instance, researchers

have observed significant dependence between the number of years of musical training [50], [139], gender [51], familiarity with songs [51], culture [50], [140], genre preference [50], and even age [50]. The latter study makes an argument for including profile information in the emotion prediction model to personalize the predictions. However, not many datasets include this type of information other than MERP.

In addition, many datasets have a cultural bias, meaning that they are annotated by people from the same culture, often simply because of the locality of the experiment or language constraints. However, it has been established that people from different countries or cultural backgrounds have different perceptions of emotion for the same music fragment [141]. For instance, [50] have raters from both the US as well as India, and see a significant difference in their annotations. HKU956 [49] go even further and include the rater's responses to a personality test: 'Ten Item Personality Measure'. In future work, one could develop datasets annotated by annotators from different cultural backgrounds, that include this information about the raters. The personalized nature of emotion ratings can cause a low inter-rater reliability, a metric of agreement between raters often calculated using Cronbach's Alpha [142].

Noisy labels When creating datasets, an additional challenge arises: it can be hard to identify the emotion perceived from music, especially when working with valence and arousal models. In fact, Russell's model [66] included a third dimension: dominance. This dimension is typically omitted because of the ambiguity in annotation. In general, categorical models, although less precise, are often easier to annotate [76]. This personal variability and ambiguity in emotion labels introduces noise in dataset labels, causing low inter-rater reliability, and making it challenging to train accurate and reliable MER models. In addition, some emotion representations are prone to more noise. For instance, the MTG-Jamendo dataset [41] offers a large number of freely assigned tags, making it suitable for exploring a wide range of emotions, but the lack of a controlled annotation process can lead to noisy data, as there may be synonyms of emotion terms.

Machine learning models may also offer useful techniques to deal with noisy labels, as discussed in the survey by [143]. These could include Noisy Graph Cleaning (NGC) [144], Joint Training with Co-Regularization (JoCoR) [145], and Robust Curriculum Learning (RoCL) [146].

Annotation interfaces When creating a static dataset with static annotations, some standard tools can be used, including PsyToolkit [147]. However, to create any larger-scale dataset, the annotations are often done through crowdsourcing services such as Amazon Mechanical Turk⁵ (e.g. for MERP [50], DEAM [37], and Solyman's dataset [28]). These services offer access to a large 'army' of annotators, which may come at the cost of accuracy. There are, however, a number of techniques that can be used to filter out some of this annotation noise. This includes limiting the annotations to Master raters (raters with a known track record that often work at a premium price) [50], by including qualification tasks to assess participants' understanding of the dimensional model [28], or by using

multiple ground truth questions that are shown to all raters [50]. Finally, inter-rater reliability can be used to filter out low-quality annotations [50]. When doing this, it is important to keep in mind personal characteristics, which may cause different people to rate music differently. Hence, inter-rater reliability should ideally be calculated by taking into account the rater's profile features.

An additional difficulty is that the amount of available interfaces for music emotion annotation is very limited. Especially when it comes to time-continuous annotations of valence and arousal, [50] released their dynamic annotation interface⁶ that hooks into Amazon mTurk. Kim et al. [21] also introduced, an annotation interface called MoodSwings, designed to record dynamic emotion labels.

Benchmarking The ImageNet competition has been instrumental in establishing a clear performance benchmark among computer vision models [148]. While there have been similar competitions for music emotion prediction (see Section III), none of these competitions ran in the last three years, indicating the lack of a current benchmark for MER systems. To establish benchmarks on individual datasets, we have to revert to individual model papers, such as [28], [37]. It is unfortunately not always clear which train/test split these systems use, making direct comparisons hard. One way to facilitate an easy overview would be to leverage Leaderboard features on popular websites such as Papers With Code⁷ and HuggingFace⁸.

While there have been some efforts toward cross-dataset comparisons, such practices are still relatively rare. A notable example is the MusAV dataset [53], which provides a benchmark for evaluating arousal-valence (AV) regression models trained on different datasets. MusAV uses relative pairwise comparisons as ground truth and enables comparative validation across models trained on diverse AV datasets. Such initiatives are promising steps toward better generalization and standardized evaluation, but they are currently exceptions rather than the rule. In general, many datasets still use different emotion representations, and cross-dataset evaluation remains challenging. Recent efforts have aimed to address this limitation: Kang et al. [124] propose a unified multitask framework to combine categorical and dimensional labels, while Liu et al. [123] use LLM-based label embeddings to align emotion annotations across datasets and enable zero-shot generalization.

Bridging between datasets that use continuous representations and categorical models is not straightforward. Several studies, such as Paltoglou and Thelwall [149], have proposed mappings between arousal/valence and categorical labels. This approach has been further utilized, for example, by Makris et al. [6] for emotion-controlled lead sheet generation. In addition, large-scale affective norm resources such as Warriner et al. [150], which provides valence, arousal, and dominance (VAD) ratings for 13,915 English lemmas, and Mohammad [151], who developed the NRC VAD Lexicon with ratings

⁶<https://github.com/dorienh/MERP>

⁷<https://paperswithcode.com/sota>

⁸<https://huggingface.co/docs/competitions/en/leaderboard>

⁵<https://www.mturk.com>

for over 20,000 words, offer valuable foundations for linking continuous and categorical representations. Such mappings could support merging continuous and categorical datasets into larger-scale resources for music emotion research.

MIDI The forgotten format in music emotion recognition. Emotion originates from many different aspects of the music, including the tonal tension [8], instrumentation and timbre [152], production quality [153], expressiveness of the performance [154], harmony [155]. The symbolic MIDI format only captures parts of these [13], which may explain the lack of models for predicting emotion from MIDI. Another reason may simply be the lack of emotion-annotated datasets (three in total). Even though MIDI is an abstraction of music, it is still a widely used format by music producers and performers and warrants its own models for emotion prediction. Even more, such datasets may enable generative music systems (which are typically trained on MIDI) to be controlled by emotion [6], [156].

Multimodal predictions Our sensory input is multimodal. As such, some of the emotion-annotated datasets enable us to look at multiple modalities. For instance, DEAP [26] offers biofeedback data, i.e. EEG recordings, and frontal face videos from participants. Similarly, HKU956 [49] offers physiological signals including heart rate, electrodermal activity, blood volume pulse, inter-beat interval, and skin temperature. SiTunes [60], on the other hand, provides physiological and environmental situation recordings collected via smart wristband devices.

These biological data can serve as the induced emotion labels, e.g. EEG signals can be translated into human emotions [157]. Increased datasets with different types of physiological signals enable researchers to focus on creating models for induced versus perceived emotion detection. This offers new avenues to use biofeedback in music emotion mediation applications through smart devices.

Sometimes other emotion-inducing modalities are present, such as video, or lyrics. In this case, it is important to consider the influence of each of these modalities. In the case of video and music, Chua et al. [51] have studied the influence of exposing participants not only to the music but also the muted video, as well as the music videos. They found that the music modality explains most of the variance in arousal values, and both music and video modalities explain the variance in valence values. Phuong et al. [158] explore the influence of using only audio features and only video features to predict emotions from movie fragments. They found that the prediction is most accurate when both modalities are used. However, when predicting from a single modality, the audio model is most accurate.

Real-time Many of the currently existing models are not implemented as an easy-to-use library, nor are they quick to run. They often require a GPU and may take several minutes to run. There are use cases, however, for real-time emotion recognition systems, as they would be able to integrate into therapeutic emotion detection systems [5], mood guidance playlist systems, as well as more commercial systems such as advertisement targeting systems.

Toward reliable and comparable MER research Despite

significant advancements, the field still lacks standardized, high-quality datasets that serve as universal benchmarks. As discussed above, most existing datasets differ widely in genre coverage, emotion representation models, annotation procedures, and data quality. This heterogeneity, combined with inconsistent use of evaluation metrics and train/test splits, often makes it difficult to compare results across studies—even when using the same dataset. For example, the MTG-Jamendo dataset has been used in numerous studies (see Table II), yet reported results vary substantially due to differing preprocessing strategies, loss functions, or model inputs. This wide variance raises questions about reproducibility and comparability in the field.

While recent initiatives like MusAV [53] and unified multi-task frameworks [124] represent important progress, the lack of standardized benchmarks still hinders progress. We argue that no single dataset can serve all research needs, especially given the subjective and culture-specific nature of emotion. Rather than searching for a one-size-fits-all dataset, we encourage the development of dataset-agnostic benchmarking tools, clearly defined splits, and model evaluation protocols. The community would benefit from centralized efforts (e.g., leaderboards, open splits, and documentation hubs) that promote transparency and make it easier to assess which models and results should be taken more seriously. Without such measures, the field risks being undermined by non-comparable results and hard-to-replicate studies.

In sum, the challenges mentioned above provide direct opportunities and future directions to further advance the exciting field of music emotion recognition.

VI. CONCLUSION

Music emotion recognition is a promising field with various practical applications. With the rise of large-language models, we have seen impressive performance in various tasks. The field of music emotion recognition, however, still seems to be lagging. Given the importance of large training datasets to facilitate the training of LLMs, we provide a comprehensive overview and discussion of the existing datasets for music emotion recognition.

We also explore current state-of-the-art models and dive into evaluation methods such as metrics as well as competitions, leaderboards, and benchmarks within the MER field. With this knowledge, we discuss the current challenges of the MER field at length and provide concrete future directions and emerging trends such as real-time systems and multimodal prediction systems.

In closing, this survey serves as a valuable resource for the MER community, by offering insights into the current state-of-the-art, as well as a discussion of challenges and inspiration for future directions.

VII. ACKNOWLEDGEMENTS

This work has received SEED funding from SUTD TL under grant number RTDS S 22 14 04 1 and SUTD SKI 2021_04_06.

REFERENCES

- [1] Meyer Leonard. Emotion and meaning in music. *Chicago: University of Chicago*, 1956.
- [2] Carl E Seashore. The psychology of music. *Music Educators Journal*, 23(4):30–33, 1937.
- [3] Kate Hovner. The affective character of the major and minor modes in music. *The American Journal of Psychology*, 47(1):103–118, 1935.
- [4] Patrik N Juslin and Daniel Västfjäll. Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and brain sciences*, 31(5):559–575, 2008.
- [5] Kat R Agres, Rebecca S Schaefer, Anja Volk, Susan van Hooren, Andre Holzapfel, Simone Dalla Bella, Meinard Müller, Martina De Witte, Dorien Herremans, Rafael Ramirez Melendez, et al. Music, computing, and health: a roadmap for the current and future roles of music technology for health care and well-being. *Music & Science*, 4:2059204321997709, 2021.
- [6] Dimos Makris, Kat R Agres, and Dorien Herremans. Generating lead sheets with affect: A novel conditional seq2seq framework. In *2021 Int. Joint Conf. on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.
- [7] Adyasha Dash and Kathleen Agres. Ai-based affective music generation systems: A review of methods and challenges. *ACM Computing Surveys*, 56(11):1–34, 2024.
- [8] Dorien Herremans and Elaine Chew. Morpheus: generating structured music with constrained patterns and tension. *IEEE Transactions on Affective Computing*, 10(4):510–523, 2017.
- [9] Kat R Agres, Adyasha Dash, and Phoebe Chua. Affectmachine-classical: a novel system for generating affective classical music. *Frontiers in Psychology*, 14:1158172, 2023.
- [10] Yi-Hsuan Yang and Homer H Chen. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):1–30, 2012.
- [11] Martin EP Seligman and Mihaly Csikszentmihalyi. *Positive psychology: An introduction.*, volume 55. American Psychological Association, 2000.
- [12] Donghong Han, Yanru Kong, Jiayi Han, and Guoren Wang. A survey of music emotion recognition. *Frontiers of Computer Science*, 16(6):166335, 2022.
- [13] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. In *Proc. ismir*, volume 86, pages 937–952, 2010.
- [14] Xinyu Yang, Yizhuo Dong, and Juan Li. Review of data features-based music emotion recognition methods. *Multimedia systems*, 24:365–389, 2018.
- [15] Mathieu Barthet, György Fazekas, and Mark Sandler. Music emotion recognition: From content-to context-based models. In *From Sounds to Music and Emotions: 9th Int. Symposium, CMMR 2012, London, UK, June 19–22, 2012, Revised Selected Papers 9*, pages 228–252. Springer, 2013.
- [16] Renato Panda, Ricardo Malheiros, and Rui Pedro Paiva. Audio features for music emotion recognition: a survey. *IEEE Transactions on Affective Computing*, 14(1):68–88, 2020.
- [17] Xingguo Jiang, Yuchao Zhang, Guojun Lin, and Ling Yu. Music emotion recognition based on deep learning: A review. *IEEE Access*, 2024.
- [18] Xiao Hu and J Stephen Downie. Exploring mood metadata: Relationships with genre, artist and usage metadata. In *ISMIR*, pages 67–72, 2007.
- [19] Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. Towards musical query-by-semantic-description using the cal500 data set. In *Proc. of the 30th annual Int. ACM SIGIR Conf. on Research and development in information retrieval*, pages 439–446, 2007.
- [20] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H Chen. A regression approach to music emotion recognition. *IEEE Transactions on audio, speech, and language processing*, 16(2):448–457, 2008.
- [21] Youngmoo E Kim, Erik M Schmidt, and Lloyd Emelle. Moodswings: A collaborative game for music mood label collection. In *ismir*, volume 8, pages 231–236, 2008.
- [22] Björn Schuller, Johannes Dorffner, and Gerhard Rigoll. Determination of nonprototypical valence and arousal in popular music: features and performances. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010:1–19, 2010.
- [23] Tuomas Eerola and Jonna K Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1):18–49, 2011.
- [24] Jacquelin A Speck, Erik M Schmidt, Brandon G Morton, and Youngmoo E Kim. A comparative study of collaborative vs. traditional musical mood annotation. In *ISMIR*, volume 104, pages 549–554, 2011.
- [25] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Ismir*, volume 2, page 10, 2011.
- [26] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deep: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.
- [27] Renato Eduardo Silva Panda, Ricardo Malheiros, Bruno Rocha, António Pedro Oliveira, and Rui Pedro Paiva. Multi-modal music emotion recognition: A new dataset, methodology and comparative analysis. In *10th Int. symposium on computer music multidisciplinary research (CMMR 2013)*, pages 570–582, 2013.
- [28] Mohammad Soleymani, Micheal N Caro, Erik M Schmidt, Cheng-Ya Sha, and Yi-Hsuan Yang. 1000 songs for emotional analysis of music. In *Proc. of the 2nd ACM Int. workshop on Crowdsourcing for multimedia*, pages 1–6, 2013.
- [29] Shuo-Yang Wang, Ju-Chiang Wang, Yi-Hsuan Yang, and Hsin-Min Wang. Towards time-varying music auto-tagging based on cal500 expansion. In *2014 IEEE Int. Conf. on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2014.
- [30] Yu-An Chen, Yi-Hsuan Yang, Ju-Chiang Wang, and Homer Chen. The amg1608 dataset for music emotion recognition. In *2015 IEEE Int. Conf. on acoustics, speech and signal processing (ICASSP)*, pages 693–697. IEEE, 2015.
- [31] Marcel Zentner, Didier Grandjean, and Klaus R Scherer. Emotions evoked by the sound of music: characterization, classification, and measurement. *Emotion*, 8(4):494, 2008.
- [32] Matevž Pesek, Gregor Strel, Alenka Kavčič, and Matija Marolt. The moodo dataset: Integrating user context with emotional and color perception of music for affective music information retrieval. *Journal of New Music Research*, 46(3):246–260, 2017.
- [33] Ricardo Malheiros, Renato Panda, Paulo JS Gomes, and Rui Pedro Paiva. 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—ECML/PKDD 2016*, 2016.
- [34] Xiao Hu and Yi-Hsuan Yang. The mood of chinese pop music: Representation and recognition. *Journal of the Association for Information Science and Technology*, 68(8):1899–1910, 2017.
- [35] Erion Çano and Maurizio Morisio. Moodylyrics: A sentiment annotated lyrics dataset. In *Proceedings of the 2017 international conference on intelligent systems, metaheuristics & swarm intelligence*, pages 118–124, 2017.
- [36] Renato Panda, Ricardo Malheiros, and Rui Pedro Paiva. Musical texture and expressivity features for music emotion recognition. In *19th Int. Society for Music Information Retrieval Conf. (ISMIR 2018)*, pages 383–391, 2018.
- [37] Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. Developing a benchmark for emotional analysis of music. *PloS one*, 12(3):e0173392, 2017.
- [38] Kejun Zhang, Hui Zhang, Simeng Li, Changyuan Yang, and Lingyun Sun. The pmemo dataset for music emotion recognition. In *Proc. of the 2018 acm on Int. Conf. on multimedia retrieval*, pages 135–142, 2018.
- [39] Steven R Livingstone and Frank A Russo. 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, 2018.
- [40] Rémi Delbouys, Romain Hennequin, Francesco Piccoli, Jimena Royo-Letelier, and Manuel Moussallam. Music mood detection based on audio and lyrics with deep neural net. *arXiv:1809.07276*, 2018.
- [41] Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. The mtg-jamendo dataset for automatic music tagging. *ICML*, 2019.
- [42] Lucas N Ferreira and Jim Whitehead. Learning to generate music with sentiment. *arXiv:2103.06125*, 2021.
- [43] Mehmet Bilal Er and Ibrahim Berkay Aydilek. Music emotion recognition by using chroma spectrogram and deep visual features. *Int. Journal of Computational Intelligence Systems*, 12(2):1622–1634, 2019.
- [44] Hsiao-Tzu Hung, Joann Ching, Seungheon Doh, Nabin Kim, Juhan Nam, and Yi-Hsuan Yang. Emopia: A multi-modal pop piano

- dataset for emotion recognition and emotion-based music generation. *arXiv:2108.01374*, 2021.
- [45] M. Velankar. Mer500 — music emotion recognition. <https://www.kaggle.com/>, June 2020. Accessed March 10, 2024.
- [46] Igor André Pegoraro Santana, Fabio Pinhelli, Juliano Donini, Leonardo Catharin, Rafael Biazus Mangolin, Valéria Delisandra Feltrim, Marcos Aurélio Domingues, et al. Music4all: A new music database and its applications. In *2020 Int. Conf. on Systems, Signals and Image Processing (IWSSIP)*, pages 399–404. IEEE, 2020.
- [47] Jianyu Fan, Yi-Hsuan Yang, Kui Dong, and Philippe Pasquier. A comparative study of western and chinese classical music based on soundscape models. In *ICASSP 2020-2020 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 521–525. IEEE, 2020.
- [48] Christopher Akiki and Manuel Burghardt. Muse: The musical sentiment dataset. *Journal of Open Humanities Data*, 7, 2021.
- [49] Xiao Hu, Fanjie Li, and Ruilun Liu. Detecting music-induced emotion based on acoustic analysis and physiological sensing: A multimodal approach. *Applied Sciences*, 12(18):9354, 2022.
- [50] En Yan Koh, Kin Wai Cheuk, Kwan Yee Heung, Kat R Agres, and Dorien Herremans. Merp: a music dataset with emotion ratings and raters' profile information. *Sensors*, 23(1):382, 2022.
- [51] Phoebe Chua, Dimos Makris, Dorien Herremans, Gemma Roig, and Kat Agres. Predicting emotion from music videos: exploring the relative contribution of visual and auditory information to affective responses. *arXiv:2202.10453*, 2022.
- [52] Eunjin Choi, Yoonjin Chung, Seolhee Lee, JongIk Jeon, Taegyun Kwon, and Juhan Nam. Ym2413-mdb: A multi-instrumental fm video game music dataset with emotion annotations. *arXiv:2211.07131*, 2022.
- [53] Dmitry Bogdanov, Xavier Lizarraga Seijas, Pablo Alonso-Jiménez, and Xavier Serra. Musav: A dataset of relative arousal-valence annotations for validation of audio models. 2022.
- [54] Ha Thi Phuong Thao, Gemma Roig, and Dorien Herremans. Emomv: Affective music-video correspondence learning datasets for classification and retrieval. *Information Fusion*, 91:64–79, 2023.
- [55] Andrew Steven Sams and Amalia Zahra. Multimodal music emotion recognition in indonesian songs based on cnn-lstm, xlnet transformers. *Bulletin of Electrical Engineering and Informatics*, 12(1):355–364, 2023.
- [56] Juan Sebastián Gómez-Cañón, Nicolás Gutiérrez-Páez, Lorenzo Portero, Alastair Porter, Estefanía Cano, Perfecto Herrera-Boyer, Aggelos Giokas, Patricia Santos, Davinia Hernández-Leo, Casper Karreman, et al. Trompa-mer: an open dataset for personalized music emotion recognition. *Journal of Intelligent Information Systems*, 60(2):549–570, 2023.
- [57] Méhania Doumbia, Maxime Renard, Laure Coudrat, and Geoffray Bonnin. Characterizing the emotional context induced by music listening and its effects on gait initiation: Exploiting physiological and biomechanical data. In *Adjunct Proc. of the 31st ACM Conf. on User Modeling, Adaptation and Personalization*, pages 182–186, 2023.
- [58] Yesid Ospitia-Medina, José Ramón Beltrán, and Sandra Baldassarri. Ensa dataset: a dataset of songs by non-superstar artists tested with an emotional analysis based on time-series. *Personal and Ubiquitous Computing*, 27(5):1909–1925, 2023.
- [59] Hannah Strauss, Julia Vigl, Peer-Ole Jacobsen, Martin Bayer, Francesca Talamini, Wolfgang Vigl, Eva Zangerle, and Marcel Zentner. The emotion-to-music mapping atlas (emma): A systematically organized online database of emotionally evocative music excerpts. *Behavior Research Methods*, pages 1–18, 2024.
- [60] Vadim Grigorev, Jiayu Li, Weizhi Ma, Zhiyu He, Min Zhang, Yiqun Liu, Ming Yan, and Ji Zhang. Situnes: A situational music recommendation dataset with physiological and psychological signals. In *Proc. of the 2024 ACM SIGIR Conf. on Human Information Interaction and Retrieval*, pages 417–421, 2024.
- [61] Pedro Lima Louro, Hugo Redinho, Ricardo Santos, Ricardo Malheiro, Renato Panda, and Rui Pedro Paiva. Merge—a bimodal dataset for static music emotion recognition. *arXiv preprint arXiv:2407.06060*, 2024.
- [62] Kinda Wu, Jiaming Wang, Jiaxing Yu, Tieyao Zhang, and Kejun Zhang. Popular hooks: A multimodal dataset of musical hooks for music understanding and generation. In *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2024.
- [63] Joanne Affolter and Martin Rohrmeier. Utilizing listener-provided tags for music emotion recognition: A data-driven approach. In *25th Int. Society for Music Information Retrieval Conf. (ISMIR 2024)*, pages 547–554, 2024.
- [64] Sida Tian, Can Zhang, Wei Yuan, Wei Tan, and Wenjie Zhu. Xmusic: Towards a generalized and controllable symbolic music generation framework. *arXiv preprint arXiv:2501.08809*, 2025.
- [65] Kate Henn. Experimental studies of the elements of expression in music. *The American journal of psychology*, 48(2):246–268, 1936.
- [66] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.
- [67] Iris Bakker, Theo Van Der Voort, Peter Vink, and Jan De Boon. Pleasure, arousal, dominance: Mehrabian and russell revisited. *Current Psychology*, 33:405–421, 2014.
- [68] Robert E Thayer. *The biopsychology of mood and arousal*. Oxford University Press, 1990.
- [69] Dmitry Bogdanov, Alastair Porter, Philip Tovstogan, and Minz Won. Mediaeval2019: Emotion and theme recognition in music using jamendo. In *Larson M, Hicks S, Constantin MG, Bischke B, Porter A, Zhao P, Lux M, Cabrera Quiros L, Calandre J, Jones G, editors. MediaEval'19, Multimedia Benchmark Workshop; 2019 Oct 27-30, Sophia Antipolis, France. Aachen: CEUR; 2019*. CEUR Workshop Proc., 2019.
- [70] Patrik N Juslin and Petri Laukka. 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, 2004.
- [71] Konstantinos Trochidis, David Sears, Diêu-Ly Trân, and Stephen McAdams. Psychophysiological measures of emotional response to romantic orchestral music and their musical and acoustic correlates. In *From Sounds to Music and Emotions: 9th Int. Symposium, CMMR 2012, London, UK, June 19-22, 2012, Revised Selected Papers 9*, pages 44–57. Springer, 2013.
- [72] Valorie N Salimpour, Mitchel Benovoy, Kevin Larcher, Alain Dagher, and Robert J Zatorre. Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. *Nature neuroscience*, 14(2):257–262, 2011.
- [73] Javier Jaimovich, Niall Coghlan, and R Benjamin Knapp. Emotion in motion: A study of music and affective response. In *From Sounds to Music and Emotions: 9th Int. Symposium, CMMR 2012, London, UK, June 19-22, 2012, Revised Selected Papers 9*, pages 19–43. Springer, 2013.
- [74] Richard A McFarland. Relationship of skin temperature changes to the emotions accompanying music. *Biofeedback and Self-regulation*, 10:255–267, 1985.
- [75] Jonghwa Kim and Elisabeth André. Emotion recognition based on physiological changes in music listening. *IEEE transactions on pattern analysis and machine intelligence*, 30(12):2067–2083, 2008.
- [76] Yading Song, Simon Dixon, Marcus T Pearce, and Andrea R Halpern. Perceived and induced emotion responses to popular music: Categorical and dimensional models. *Music Perception: An Interdisciplinary Journal*, 33(4):472–492, 2016.
- [77] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.
- [78] Jianyu Fan, Miles Thorogood, and Philippe Pasquier. Emo-soundscapes: A dataset for soundscape emotion recognition. In *2017 Seventh international conference on affective computing and intelligent interaction (ACII)*, pages 196–201. IEEE, 2017.
- [79] Yi-Hsuan Yang and Homer H Chen. Ranking-based emotion recognition for music organization and retrieval. *IEEE Transactions on audio, speech, and language processing*, 19(4):762–774, 2010.
- [80] Jianyu Fan, Kivanç Tatar, Miles Thorogood, and Philippe Pasquier. Ranking-based emotion recognition for experimental music. In *ISMIR*, volume 2017, pages 368–375, 2017.
- [81] Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4, 2009.
- [82] Kin Wai Cheuk, Yin-Jyun Luo, BT Balamurali, Gemma Roig, and Dorien Herremans. Regression-based music emotion prediction using triplet neural networks. In *2020 Int. joint Conf. on neural networks (ijcn)*, pages 1–7. IEEE, 2020.
- [83] Srividya Tirunellai Rajamani, Kumar Rajamani, and Björn Schuller. Emotion and theme recognition in music using attention-based methods. 2020.
- [84] Maximilian Mayerl, Michael Vötter, Andreas Peintner, Günther Specht, and Eva Zangerle. Recognizing song mood and theme: Clustering-based ensembles. In *MediaEval*, 2021.
- [85] Hao Hao Tan. Semi-supervised music emotion recognition using noisy student training and harmonic pitch class profiles. *arXiv:2112.00702*, 2021.

- [86] Vincent Bour. Frequency dependent convolutions for music tagging. In *MediaEval*, 2021.
- [87] Hardik Sharma, Shelly Gupta, Yukti Sharma, and Archana Purwar. A new model for emotion prediction in music. In *2020 6th International Conference on Signal Processing and Communication (ICSC)*, pages 156–161. IEEE, 2020.
- [88] Na He and Sam Ferguson. Multi-view neural networks for raw audio-based music emotion recognition. In *2020 IEEE International Symposium on Multimedia (ISM)*, pages 168–172. IEEE, 2020.
- [89] Sangeetha Rajesh and NJ Nalini. Musical instrument emotion recognition using deep recurrent neural network. *Procedia Computer Science*, 167:16–25, 2020.
- [90] Changfeng Chen and Qiang Li. A multimodal music emotion classification method based on multifeature combined network classifier. *Mathematical Problems in Engineering*, 2020(1):4606027, 2020.
- [91] Sanga Chaki, Pranjal Doshi, Priyadarshi Patnaik, and Sourangshu Bhattacharya. Attentive rnns for continuous-time emotion prediction in music clips. In *AffCon@ AAAI*, pages 36–46, 2020.
- [92] Jacopo De Berardinis, Angelo Cangelosi, and Eduardo Coutinho. The multiple voices of musical emotions: source separation for improving music emotion recognition models and their interpretability. In *ISMIR*, pages 310–317, 2020.
- [93] Mladen Russo, Luka Kraljević, Maja Stella, and Marjan Sikora. Cochleogram-based approach for detecting perceived emotions in music. *Information Processing & Management*, 57(5):102270, 2020.
- [94] Rajib Sarkar, Sombuddha Choudhury, Saikat Dutta, Aneek Roy, and Sanjoy Kumar Saha. Recognition of emotion in music based on deep convolutional neural network. *Multimedia Tools and Applications*, 79(1):765–783, 2020.
- [95] Dillon Knox, Timothy Greer, Benjamin Ma, Emily Kuo, Krishna Somandepalli, and Shrikanth Narayanan. Mediaeval 2020 emotion and theme recognition in music task: Loss function approaches for multi-label music tagging. In *MediaEval*, 2020.
- [96] Yin Yu. Research on music emotion classification based on cnn-lstm network. In *2021 5th Asian Conference on Artificial Intelligence Technology (ACAIT)*, pages 473–476. IEEE, 2021.
- [97] Serhat Hizlisoy, Serdar Yildirim, and Zekeriya Tufekci. Music emotion recognition using convolutional long short term memory deep neural networks. *Engineering Science and Technology, an International Journal*, 24(3):760–767, 2021.
- [98] Phu-Thinh Pham, Minh-Hieu Huynh, Hai-Dang Nguyen, and Minh-Triet Tran. Selab-hcmus at mediaeval 2021: Music theme and emotion classification with co-teaching training strategy. In *MediaEval*, 2021.
- [99] Jacek Grekow. Music emotion recognition using recurrent neural networks and pretrained models. *Journal of Intelligent Information Systems*, 57(3):531–546, 2021.
- [100] Yudhik Agrawal, Ramaguru Guru Ravi Shanker, and Vinoo Alluri. Transformer-based approach towards music emotion recognition from lyrics. In *European Conf. on information retrieval*, pages 167–175. Springer, 2021.
- [101] Shreyan Chowdhury, Verena Praher, and Gerhard Widmer. Tracing back music emotion predictions to sound sources and intuitive perceptual qualities. *arXiv:2106.07787*, 2021.
- [102] I-Sheng Huang, Yu-Hsuan Lu, Muhammad Shafiq, Asif Ali Laghari, and Rahul Yadav. A generative adversarial network model based on intelligent data analytics for music emotion recognition under iot. *Mobile Information Systems*, 2021:1–8, 2021.
- [103] Yagya Raj Pandeya and Joonwhao Lee. Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimedia Tools and Applications*, 80(2):2887–2905, 2021.
- [104] Darryl Griffiths, Stuart Cunningham, Jonathan Weinel, and Richard Picking. A multi-genre model for music emotion recognition using linear regressors. *Journal of New Music Research*, 50(4):355–372, 2021.
- [105] Yu Xia, Fumei Xu, et al. Study on music emotion recognition based on the machine learning model clustering algorithm. *Mathematical Problems in Engineering*, 2022, 2022.
- [106] Jibao Qiu, CL Chen, and Tong Zhang. A novel multi-task learning method for symbolic music emotion recognition. *arXiv:2201.05782*, 2022.
- [107] Le Cai, Sam Ferguson, Haiyan Lu, and Gengfa Fang. Feature selection approaches for optimising music emotion recognition methods. *arXiv:2212.13369*, 2022.
- [108] Rafael Alexandre Portugal Matos. Merge lyrics: Music emotion recognition next generation–lyrics classification with deep learning. Master’s thesis, 2022.
- [109] Na He and Sam Ferguson. Music emotion recognition based on segment-level two-stage learning. *International Journal of Multimedia Information Retrieval*, 11(3):383–394, 2022.
- [110] Yesid Ospitia Medina, José Ramón Beltrán, and Sandra Baldassarri. Emotional classification of music using neural networks with the mediaeval dataset. *Personal and Ubiquitous Computing*, 26(4):1237–1249, 2022.
- [111] Tibor Krosl, Yana Nikolova, and Ninell Oldenburg. Multi-modality in music: Predicting emotion in music from high-level audio features and lyrics. *arXiv:2302.13321*, 2023.
- [112] Xiao Han, Fuyang Chen, and Junrong Ban. Music emotion recognition based on a neural network with an inception-gru residual structure. *Electronics*, 12(4):978, 2023.
- [113] Yingjin Song and Daniel Beck. Modeling emotion dynamics in song lyrics with state space models. *Transactions of the Association for Computational Linguistics*, 11:157–175, 2023.
- [114] R Shanker, B Manikanta Gupta, BV Koushik, and Vinoo Alluri. Tollywood emotions: Annotation of valence-arousal in telugu song lyrics. *arXiv:2303.09364*, 2023.
- [115] Meixian Zhang, Yonghua Zhu, Wenjun Zhang, Yunwen Zhu, and Tianyu Feng. Modularized composite attention network for continuous music emotion recognition. *Multimedia Tools and Applications*, 82(5):7319–7341, 2023.
- [116] Maria Jose Lucia-Mulas, Pablo Revuelta-Sanz, Belen Ruiz-Mezcua, and Israel Gonzalez-Carrasco. Automatic music emotion classification model for movie soundtrack subtitling based on neuroscientific premises. *Applied Intelligence*, 53(22):27096–27109, 2023.
- [117] Sujeesha Ajithakumari Suresh Kumar and Rajeev Rajan. Transformer-based automatic music mood classification using multi-modal framework. *Journal of Computer Science & Technology*, 23, 2023.
- [118] Xinyu Chang, Xiangyu Zhang, Haoruo Zhang, and Yulu Ran. Music emotion prediction using recurrent neural networks. *arXiv preprint arXiv:2405.06747*, 2024.
- [119] Jingyi Wang, Alireza Sharifi, Thippa Reddy Gadekallu, and Achyut Shankar. Mmd-mii model: a multilayered analysis and multimodal integration interaction approach revolutionizing music emotion classification. *International Journal of Computational Intelligence Systems*, 17(1):99, 2024.
- [120] Love Jhoye Moreno Raboy and Attaphongse Taparugssanagorn. Verse1-chorus-verse2 structure: A stacked ensemble approach for enhanced music emotion recognition. *Applied Sciences*, 14(13):5761, 2024.
- [121] Xiao Han, Fuyang Chen, and Junrong Ban. A gai-based multi-scale convolution and attention mechanism model for music emotion recognition and recommendation from physiological data. *Applied Soft Computing*, 164:112034, 2024.
- [122] Jiajia Li, Samaneh Soradi-Zeid, Amin Yousefpour, and Daohua Pan. Improved differential evolution algorithm based convolutional neural network for emotional analysis of music data. *Applied Soft Computing*, 153:111262, 2024.
- [123] Renhang Liu, Abhinaba Roy, and Dorien Herremans. Leveraging llm embeddings for cross dataset label alignment and zero shot music emotion prediction. *arXiv preprint arXiv:2410.11522*, 2024.
- [124] Jaeyong Kang and Dorien Herremans. Towards unified music emotion recognition across dimensional and categorical models. *arXiv preprint arXiv:2502.03979*, 2025.
- [125] Yazhong Feng, Yueteng Zhuang, and Yunhe Pan. Music information retrieval by detecting mood via computational media aesthetics. In *Proc. IEEE/WIC Int. Conf. on web intelligence (WI 2003)*, pages 235–241. IEEE, 2003.
- [126] Lie Lu, Dan Liu, and Hong-Jiang Zhang. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on audio, speech, and language processing*, 14(1):5–18, 2005.
- [127] Tri-Nhan Do, Minh-Tri Nguyen, Hai-Dang Nguyen, Minh-Triet Tran, and Xuan-Nam Cao. Hcmus at mediaeval 2020: Emotion classification using wavenet feature with specaugment and efficientnet. In *MediaEval*, 2020.
- [128] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, et al. Mert: Acoustic music understanding model with large-scale self-supervised training. *arXiv preprint arXiv:2306.00107*, 2023.
- [129] Matthew C McCallum, Filip Korzeniowski, Sergio Oramas, Fabien Gouyon, and Andreas F Ehmann. Supervised and unsupervised learning of audio representations for music understanding. *arXiv preprint arXiv:2210.03799*, 2022.

- [130] Pablo Alonso-Jiménez, Xavier Serra, and Dmitry Bogdanov. Efficient supervised training of audio transformers for music representation learning. *arXiv preprint arXiv:2309.16418*, 2023.
- [131] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proc. of the 18th ACM Int. Conf. on Multimedia*, pages 1459–1462, 2010.
- [132] Yuan-Yuan Shi, Xuan Zhu, Hyoong-Gook Kim, and Ki-Wan Eom. A tempo feature via modulation spectrum analysis and its application to music emotion classification. In *2006 IEEE Int. Conf. on Multimedia and Expo*, pages 1085–1088. IEEE, 2006.
- [133] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv:1301.3781*, 2013.
- [134] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of the IEEE Conf. on computer vision and pattern recognition*, pages 770–778, 2016.
- [135] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proc. of the IEEE Int. Conf. on computer vision*, pages 2758–2766, 2015.
- [136] Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. Mustango: Toward controllable text-to-music generation. *Proc. of NAACL*, 2024.
- [137] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv:2104.01778*, 2021.
- [138] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36, 2024.
- [139] César F Lima and São Luís Castro. Emotion recognition in music changes across the adult life span. *Cogn. Emot.*, 25(4):585–598, 2011.
- [140] Xin Wang, Yujia Wei, Lena Heng, and Stephen McAdams. A cross-cultural analysis of the influence of timbre on affect perception in western classical music and chinese music traditions. *Frontiers in Psychology*, 12:732865, 2021.
- [141] Harin Lee, Frank Hoeger, Marc Schoenwiesner, Minsu Park, and Nori Jacoby. Cross-cultural mood perception in pop songs and its alignment with mood detection algorithms. *arXiv:2108.00768*, 2021.
- [142] J Martin Bland and Douglas G Altman. Statistics notes: Cronbach's alpha. *Bmj*, 314(7080):572, 1997.
- [143] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 2022.
- [144] Zhi-Fan Wu, Tong Wei, Jianwen Jiang, Chaojie Mao, Mingqian Tang, and Yu-Feng Li. Ngc: A unified framework for learning with open-world noisy data. In *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, pages 62–71, 2021.
- [145] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proc. of the IEEE/CVF Conf. on computer vision and pattern recognition*, pages 13726–13735, 2020.
- [146] Tianyi Zhou, Shengjie Wang, and Jeff Bilmes. Robust curriculum learning: from clean label detection to noisy label self-correction. In *Int. Conf. on Learning Representations*, 2020.
- [147] Gijsbert Stoet. Psytoolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, 44(1):24–31, 2017.
- [148] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conf. on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [149] Georgios Paltoglou and Michael Thelwall. Seeing stars of valence and arousal in blog posts. *IEEE Transactions on Affective Computing*, 4(1):116–123, 2012.
- [150] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45:1191–1207, 2013.
- [151] Saif Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 174–184, 2018.
- [152] Stephen McAdams and Bruno L Giordano. The perception of musical timbre. 2014.
- [153] David Ronan, Joshua D Reiss, and Hatice Gunes. An empirical approach to the relationship between emotion and music production quality. *arXiv:1803.11154*, 2018.
- [154] Patrik N Juslin and Renee Timmers. Expression and communication of emotion in music performance. *Handbook of music and emotion: Theory, research, applications*, pages 453–489, 2010.
- [155] Morwaread Mary Farbood. *A quantitative, parametric model of musical tension*. PhD thesis, Massachusetts Institute of Technology, 2006.
- [156] Hao Hao Tan and Dorien Herremans. Music fadernets: Controllable music generation based on high-level features via low-level feature modelling. *Proc. of ISMIR*, 2020.
- [157] Md Mustafizur Rahman, Ajay Krishno Sarkar, Md Amzad Hossain, Md Selim Hossain, Md Rabiu Islam, Md Biplob Hossain, Julian MW Quinn, and Mohammad Ali Moni. Recognition of human emotions using eeg signals: A review. *Computers in Biology and Medicine*, 136:104696, 2021.
- [158] Ha Thi Phuong Thao, BT Balamurali, Dorien Herremans, and Gemma Roig. Attendaffectnet: Self-attention based networks for predicting affective responses from movies. In *2020 25th Int. Conf. on Pattern Recognition (ICPR)*, pages 8719–8726. IEEE, 2021.



Jaeyong Kang is a Postdoctoral Research Fellow at the Singapore University of Technology and Design (SUTD). He earned his Ph.D. in Electrical Engineering and Computer Science from the Gwangju Institute of Science and Technology (GIST), South Korea, in 2017. From 2018 to 2019, he served as a Research Scientist at the Biomedical Research Institute, Seoul National University Hospital (SNUH). His research interests span a diverse range of fields, including music generation, affective computing, deep learning, computer vision, natural language processing, agent-based systems, and recommender systems.



Dorien Herremans is an Associate Professor at Singapore University of Technology and Design (SUTD), where she leads the Audio, Music, and AI (AMAAI) Lab. Her research focuses on developing cutting-edge AI technologies for multimodal applications, with a focus on generative models and affective computing. Before joining SUTD, she was a Marie Skłodowska-Curie Postdoctoral Fellow at the Centre for Digital Music at Queen Mary University of London. Prof. Dorien Herremans has been a pioneer in the music technology field, publishing her first generative music model 20 years ago.

She was also nominated on the Singapore 100 Women in Tech list in 2021, and shortlisted as one of the top 30 SAIL Award (Super AI Leader) Finalists in 2024 at the World AI Conference.