applied sciences

MDPI

Article

Detecting Music-Induced Emotion Based on Acoustic Analysis and Physiological Sensing: A Multimodal Approach

Xiao Hu [1,*], Fanjie Li [2] and Ruilun Liu [2]

[1] University of Hong Kong Shenzhen Institute of Research and Innovation, Shenzhen Virtual University Park, Shenzhen 518001, China
[2] Faculty of Education, University of Hong Kong, Pokfulam Road, Hong Kong SAR, China
* Correspondence: xiaoxhu@hku.hk

Abstract: The subjectivity of listeners' emotional responses to music is at the crux of optimizing emotion-aware music recommendation. To address this challenge, we constructed a new multimodal dataset ("HKU956") with aligned peripheral physiological signals (i.e., heart rate, skin conductance, blood volume pulse, skin temperature) and self-reported emotion collected from 30 participants, as well as original audio of 956 music pieces listened to by the participants. A comprehensive set of features was extracted from physiological signals using methods in physiological computing. This study then compared performances of three feature sets (i.e., acoustic, physiological, and combined) on the task of classifying music-induced emotion. Moreover, the classifiers were also trained on subgroups of users with different Big-Five personality traits for further customized modeling. The results reveal that (1) physiological features contribute to improving performance on valence classification with statistical significance; (2) classification models built for users in different personality groups could sometimes further improve arousal prediction; and (3) the multimodal classifier outperformed single-modality ones on valence classification for most user groups. This study contributes to designing music retrieval systems which incorporate user physiological data and model listeners' emotional responses to music in a customized manner.

Keywords: multimodal recognition; music retrieval and generation; physiological measures; sound and music computing; customization

check for updates

Citation: Hu, X.; Li, F.; Liu, R. Detecting Music-Induced Emotion Based on Acoustic Analysis and Physiological Sensing: A Multimodal Approach. *Appl. Sci.* **2022**, *12*, 9354. https://doi.org/10.3390/app12189354

Academic Editors: Goffredo Haus and Adriano Baratè

Received: 17 August 2022
Accepted: 16 September 2022
Published: 18 September 2022

1. Introduction

Music discovery based on emotion has been a universal behavior in our everyday life and has been widely exploited for supporting emotion regulation and music therapy [1]. Geared toward this music information retrieval (MIR) scenario, music emotion recognition (MER) has become an important task in the MIR field, and fruitful results have been achieved through modeling music emotion in multiple feature domains (e.g., acoustic signals, lyrics, social tags, and album images) [2].

Notwithstanding the remarkable progress achieved by MER research to date, the subjectivity of affective perception of music is still at the crux of optimizing user experience for emotion-aware MIR systems. As we frequently observe, a music piece pleasant for some people could be annoying for some others. Modeling listeners' emotional responses to music thus becomes crucial for emotion-aware music recommendation.

Individual differences in affective perception of music have been studied intensively in music psychology. It has been found that listeners' emotional response to music is not only related to music characteristics but could also be influenced by various user-centric factors, including personality traits [3], music preference [4], and listening environment [5]. This poses challenges to the content-based MER approach, which primarily relies on acoustic modality. Fortunately, the rapid development of wearable technology has enabled us to model listeners' emotional responses to music via fine-grained and time-sensitive user data such as physiological signals. With physiological signals collected via wireless and low-profile devices (e.g., wristband) during music listening, it would be possible to monitor and measure listeners' emotional responses to music in an objective and unobtrusive manner based on which music recommendation could be more customized and context-aware.

To this end, this study aimed to explore the potentials for incorporating physiological sensing for detecting music-induced emotion. Particularly, we focused on exploiting peripheral physiological signals such as heart rate, blood volume pulse, skin conductance, and skin temperature as these signals were reported to be associated with people's emotional status [6] and can be collected by portable wearable and unobtrusive devices. The research questions of this study could be summarized as:

RQ1: Do features extracted from peripheral physiological signals from listeners improve the performance of music-induced emotion classification compared to acoustic features extracted from music content?

Furthermore, motivated by studies on the role of personality traits in affective perception of music in the psychology literature [7], this study further investigated:

RQ2: Does building multimodal models on subgroups of users with similar personality traits improve classification performance of music-induced emotion?

To answer these research questions, a user experiment was conducted to record users' interactions with a novel MIR system, in which participants' self-reported emotional status and a series of peripheral physiological signals (e.g., heart rate, skin conductance, skin temperature) was collected by a research-grade wristband. These data were then aligned with 956 music pieces traced from participants' music listening history. In response to RQ1, we trained classification models on three distinct feature sets, including acoustic features extracted from music content, physiological features extracted from listeners' physiological signals, and the combination of the two, and compared the performances across these sources of different modalities. Moreover, classifiers were also trained on subgroups of users with similar Big-Five personality traits for further customized modeling.

Contributions of this study are four-fold. First, there are few datasets in MER involving listeners' physiological responses to music pieces and the original audio of the music, as well as emotion labels reported by the listeners. By constructing and sharing a new dataset of 956 sets of aligned physiological signals, (copyright-free) music audio and user-reported emotion, this study provides an open testbed for combining the two modalities for MER. Second, as physiological signals have not yet been thoroughly exploited in the domain of MIR, this study explores methods of physiological signal processing and extracts a fairly comprehensive set of physiological features, including those specific to electrodermal activity (EDA) and heart rate variability (HRV), further bridging the literature of MIR and physiological computing. Third, through comparing MER performances of features in combined and individual modalities, this study helps verify the possibility of designing MIR systems that can make use of physiological data and model listeners' emotional responses to music. Last but not least, through connecting music and human emotion at the physiological level, deliverables of this research can be applicable to related fields such as music perception, music therapy [8], human-computer interaction, and well-being studies [9].

2. Related Work

2.1. Music and Emotion

Exploiting the power of music in mood modulation has become one of the primary reasons for people's engagement with music [3], which has motivated an increasing body of research geared toward the design and implementation of emotion-aware MIR systems. Music emotion recognition (MER) is the underlying technique for designing emotion-aware music retrieval systems and has been an important part of community-wide evaluation campaigns such as Music Information Retrieval Evaluation Exchange (MIREX) since 2007 [10] and MediaEval since 2013 [11]. In MER research, there are two distinct kinds of music emotion to be automatically recognized based on which two research paradigms have been developed. One is *perceived* emotion, the other is *induced* emotion [12].

As stressed by Juslin et al. [13], there is an intrinsic distinction between the emotion expressed, the emotion perceived, and the emotion induced by music. Specifically, the expressed emotion focuses on the expressive intention of the composer or performer, while the perceived emotion refers to how listeners perceive the emotion conveyed by the music pieces based on the musical features [13]. Furthermore, given the possibility that a listener might perceive what a music piece tries to convey but not necessarily feel that emotion, Juslin et al. [13] further conceptualizes the induced emotion, which is the actual emotional experience of listeners. As the expressive intention of the composer/performer is beyond the scope of MER task, emotion-ground truth generally refers to perceived or induced emotions [12].

From the perspective of perceived emotion, a wide range of music features have been reported to be related to the emotion a music piece conveys, including rhythm and tempo, loudness, mode, pitch and consonance, tone attacks and decays, timbre, and melody [14]. This psychological basis of predicting music emotion from acoustic modality has motivated a considerable number of studies in the MIR field, which is today generally referred to as the content-based MER approach. Among the multiple feature domains (e.g., acoustic signals [15], lyrics [16]) exploited in these studies, the acoustic signal domain is the earliest and most fundamental modality and has been widely used in online digital music services (e.g., Spotify).

Nevertheless, as indicated by the distinction between perceived and induced emotion, despite the implicit music-centered assumption in the content-based MER research, people's emotional experience of music is subjective and temporal, which may not solely depend on characteristics of the music piece. Emotion induced by music is also influenced by personal and situational factors, such as personality traits, music preference, and listening environment [3–5]. Hence, toward the ultimate goal of optimizing user experience in emotion-aware music retrieval, it would be important to incorporate user-related modalities into current MER framework.

2.2. Recognition of Music-Induced Emotion

Existing MER research has been dominated by the music-centered paradigm, focusing on perceived emotion [8,9], while the MER task on music-induced emotion started later yet is gaining increasing interest [17–19]. The Emotify dataset constructed by Aljanaki et al. [17] is one of the few datasets with music-induced emotion annotated by human annotators. It has since been frequently used in subsequent studies focusing on modeling emotion induced by music listening through content-based analysis (e.g., [18,19]). In these studies, different sets of acoustic features that capture music characteristics in loudness, rhythm, timbre, pitch, and harmony have been exploited [17–19]. As pointed out in [17], results in these studies indicate that the content-based modeling of induced emotion seems to be more feasible for some emotion categories (e.g., tenderness, joyful activation) than others. Although later studies (e.g., [19]) further improved the prediction performance on other emotion categories such as calmness and power, there is still room for improvement on emotions such as amazement and sadness. One possible direction of further research suggested by [17] was to incorporate contextual information for modeling emotion categories.

Physiological signals, as a user- and context-related information source, have gained increasing attention for recognizing emotion induced by music in recent years, where both brain signals [20–23] and peripheral physiological signals such as electrocardiography (ECG) [24] were exploited for this research topic. For example, refs. [21–23] conducted comprehensive laboratory experiments to collect participants' electroencephalography (EEG) under music stimuli and self-reported emotion states. Statistical regression models [21] and several machine learning algorithms [22] were modeled to predict induced emotion. In particular, apart from the features extracted from EEG signals, ref. [21] leveraged music acoustic features into the models and resulted in significantly better prediction on the induced valence and arousal. In addition, multimodal views of people's emotional response to music, which comprehensively considered acoustic characteristics of music and physiological responses of listeners, have also contributed to unravelling the underlying mechanism of individuals' affective perception and emotional experience of music [25,26]. For instance, refs. [25,26] showed that loudness-, rhythm-, and harmony-related acoustic features were informative for predicting people's emotional responses to music, and both studies revealed associations between acoustic features and physiological measures. These findings provide empirical evidence for understanding the interplay between the psychoacoustic characteristics of music and listeners' emotional and physiological responses to music.

2.3. Physiology-Enhanced MIR Systems

As physiological changes are an inherent part of emotion [6], with the rapid development of wearable technology, physiological signals are of great potential for enhancing the state-of-art of MER. However, studies exploiting physiological signals for MER are still scarce, with the following being pioneering exceptions.

Healey et al.'s Affective DJ [27] was an early study on personalizing music selection for individual listeners based on their physiological responses to music. Specifically, the study inferred users' emotional state from their galvanic skin response (GSR) and automatically generated a relaxing or energizing playlist based on the inferred arousal level of the user. Their evaluation revealed a significant correlation between participants' arousal levels and skin conductance. Following up, Janssen et al. proposed another affective music player named AMP and probed the relationship between skin temperature and valence [28] as well as that between skin conductance and arousal [29]. Validation of the AMP found that decreasing skin temperature was correlated with more positive emotion induced by music. A few studies also tried to incorporate other types of physiological signals into MIR systems as well such as electrocardiogram (ECG) [30], photoplethysmography (PPG) [31], and electroencephalography (EEG) [32]. For instance, Chiu et al. [30] designed a music selection system which used heart rate variability (HRV) to infer arousal status of users such as sleep, boredom, anxiety, and panic and performed music selection based on a set of prespecified rules to maintain users' arousal at a moderate level.

More recently, Kim et al. [31] proposed a more sophisticated framework for emotion-aware music recommendation where features extracted from listeners' galvanic skin response (GSR) and photoplethysmography (PPG) signals were used to predict their emotional status (i.e., positive vs. negative valence; high vs. low arousal). Hu et al. [33] also explored the relationships between music-induced emotion and physiological signals of listeners and found some features extracted from heart rate (HR) and electrodermal activity (EDA) differed significantly across emotion categories.

Notwithstanding the contributions of these existing studies, the investigation on exploiting physiological signals for emotion-aware MIR is still limited. One limiting factor is the lack of publicly available datasets. To date, the most widely used dataset on this topic is DEAP [34], which contains EEG and physiological signals of 32 participants recorded while they watched 40 pieces of music video. However, the stimuli in DEAP were not exactly music but music videos which included a large amount of dynamic visual information. It is not possible to separate the effect of auditory input using DEAP. The PMEmo dataset [35], on the other hand, used music as the stimuli, specifically, 794 music chorus clips annotated by 457 subjects. While PMEmo contains moment-to-moment emotion annotations, it only includes one type of physiological signals, electrodermal activity (EDA). More recently, a new dataset named MUSEC [36] was constructed with EEG signals of 20 listeners when they listened to 220 pieces of music. While the study [36] collected listeners' ratings on valence and arousal, these emotion labels were not released with the dataset. Thus, a dataset with multiple physiological signals, corresponding music stimuli, and emotion labels is much in need.

Another research gap lies in insufficient work on features extracted from physiological signals. Most studies exploited the time-series nature of physiological signals and extracted features in the time-domain (e.g., means and standard deviations) and frequency-domain (e.g., power spectrum) [37]. Other studies employed deep neural network machine learning methods to automatically generated features [31], which suffers from the "black box problem" in interpreting meanings of the features. Very few studies leveraged the theoretical and empirical findings in physiological computing to extract features that can be explained by knowledge in human physiology.

Moreover, it has been found that the relationship between users' physiological signals and emotional responses to stimuli could be affected by individual differences such as personality traits [7]. However, to the best of our knowledge, few studies have empirically probed this question. The study presented in [33] is an exception in which classification of music-induced emotion was conducted for subgroups of users with different personality traits, but it only explored physiological features without considering acoustic features extracted from music. This study aims to bridge these gaps.

3. The Dataset Built from a User Experiment

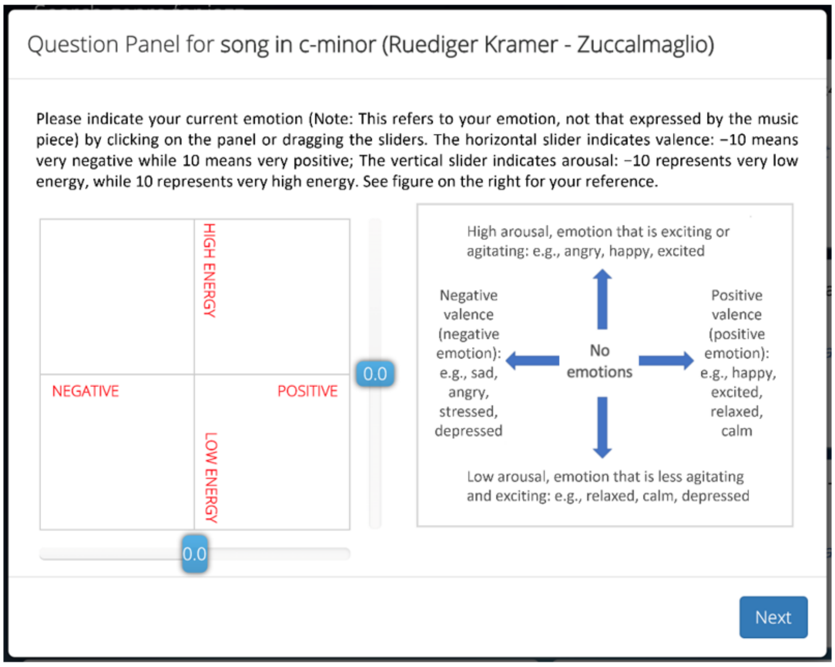
To answer the research questions, a dataset was collected through a user experiment that simulated a real-life music discovery scenario. This section describes the experimental setup and then summarizes the dataset collected.

3.1. The User Experiment Setup

This was a controlled experiment in which participants were recruited to conduct music searching and listening with a web-based music retrieval system in a quiet room. To encourage participants to interact with music, they were asked to create a playlist for the songs they liked during the experiment. Participants' peripheral physiological signals during music listening were recorded by a research-grade wristband. Moreover, we also collected their self-reported emotional responses to music via short pop-up surveys, which were taken as the ground truth labels for the classification experiment.

The experiment was conducted with a novel web-based music retrieval system, Moody (version 4), which supported music searching and browsing, online music streaming, playlist management, and interactive online surveys. In the meantime, users' interactions with Moody (e.g., play/pause music, skip a track) were logged as timestamped user events. The music collection hosted in the system is the one used in the MIREX Grand Challenge on User Experience (GC14UX) [38], which consists of 10K full tracks obtained from a free music service, Jamendo, under the Creative Commons (CC-BY) license. The collection also contains album cover images and metadata of the tracks originally obtained from Jamendo (e.g., title, album, artist, genre), which are displayed in the user interface to facilitate searching and browsing. Similar to many streaming music services, the album cover and metadata were displayed when a song was played (see [33] for a screenshot of the interface).

A pop-up survey subsystem was implemented in Moody to collect participants' self-reported emotional responses to music they listened to, which serve as ground truth for training and evaluating MER classifiers. Specifically, for each music piece listened to for more than 30 s, a question would be prompted to solicit participants' emotion induced by the song (Figure 1). The question asked participants to give scores of arousal and valence levels [39] on a continuum of -10 (very low energy, very negative) to 10 (very high energy, very positive).



Question Panel for song in c-minor (Ruediger Kramer - Zuccalmaglio)

Please indicate your current emotion (Note: This refers to your emotion, not that expressed by the music piece) by clicking on the panel or dragging the sliders. The horizontal slider indicates valence: −10 means very negative while 10 means very positive; The vertical slider indicates arousal: −10 represents very low energy, while 10 represents very high energy. See figure on the right for your reference.

HIGH ENERGY
NEGATIVE
POSITIVE
LOW ENERGY
0.0
0.0
High arousal, emotion that is exciting or agitating: e.g., angry, happy, excited
Negative valence (negative emotion): e.g., sad, angry, stressed, depressed
No emotions
Positive valence (positive emotion): e.g., happy, excited, relaxed, calm
Low arousal, emotion that is less agitating and exciting: e.g., relaxed, calm, depressed
Next

Figure 1. Pop-up question on users' emotion in the Moody system.

To probe the role of personal characteristics in predicting music-induced emotion, a pre-experiment questionnaire was administered to gather information on participants' demographics and their personality traits based on the Big-Five model (i.e., openness to experiences, conscientiousness, extraversion, agreeableness, emotional stability) measured by the Ten Item Personality Inventory (TIPI) [40].

The instructions of the Moody system and search task were subsequently presented to participants in detail. Following that, participants were asked to interact with the Moody system for no less than 40 minutes, looking for at least 10 songs they liked and adding them to their personal playlists. They were encouraged to explore music in different genres for a more diverse listening experience. During music listening, the Empatica E4 wristband, one of the few research-grade wearable devices, was used to measure participants' physiological signals. This device supports real-time data streaming and provides a Web API for accessing raw physiological data. It has been employed in emotion-related studies in neuroscience, human-computer interaction, etc., with high reliability (e.g., [41]). For signal stabilization and baseline acquisition, the wristband was mounted on each participant's wrist two minutes before the pre-experiment questionnaire.

After completing the tasks of the music search and listening, participants were asked to fill out a post-experiment questionnaire concerning their satisfaction with the music collected, emotional states at that moment, and their feedback on the Moody system and the experiment process. Informed consent forms were signed at the beginning of the experiment session, and a nominal remuneration was paid upon the completion of the experiment to compensate participants' time. This study was approved by the Human Research Ethics Committee (HREC) of the University of Hong Kong (HREC reference number: EA1802092).

3.2. Collected Data

Thirty undergraduate and postgraduate students (18 females) in a comprehensive university in Hong Kong were recruited as the participants of this experiment. According to participants' responses to the pre-experiment survey, they showed a diverse background in musical training and a relatively high frequency of music listening in everyday life (from a weekly to a daily basis).

To facilitate subgroup modeling, we grouped the participants on the five personality trait dimensions (i.e., high versus low) according to the TIPI norms provided by Gosling et al. [40]. Specifically, participants who rated lower than the TIPI normative value in a trait dimension were grouped into the “Low” category of that dimension, while those rated higher were grouped into the “High” category. For example, people who scored lower in “openness” would tend to avoid being exposed to new experiences, while people scoring high in “conscientiousness” are usually self-disciplined and prefer following plans to spontaneous actions. The number of participants in each group of each dimension is shown in Table 1.

Table 1. Specification of participant grouping.

Personality Traits	Number of Participants		TIPI Norms
	Low	High	
Openness	18	12	5.38
Conscientiousness	17	13	5.40
Extraversion	15	15	4.44
Agreeableness	24	6	5.23
Emotional stability	18	12	4.83

Note: The TIPI items are measured on 7-point Likert scales.

Overall, we collected physiological signals as well as arousal and valence ratings for 956 music pieces (592 unique ones) listened to by the participants. According to the system logs, each song was played for approximately 50 s on average.

Table 2 summarizes the data collected from the user experiment. Particularly, participants’ skin temperature was measured as degrees on the Celsius (°C) scale. Their skin conductance was collected by the EDA sensor in microsiemens (μS), and blood volume pulse (BVP), by a photoplethysmograph (PPG) sensor embedded in the Empatica E4 wristband. In addition, some secondary signals were also extracted from the BVP signal, including heart rate (HR) sampled at 1 Hz and interbeat interval (IBI), which supports heart rate variability (HRV) analysis. The raw physiological data were subsequently aligned with the starting and ending time of each music piece and split into chunks corresponding to the timestamped listening records stored in the Moody system logs.

Table 2. Description of collected data.

Category	Source	Description
Audio samples	Jamendo	MP3 audio files with accompanied metadata.
Physiological signals	Empatica E4 wristband	Blood volume pulse (BVP), heart rate (HR), interbeat interval (IBI), electrodermal activity (EDA), and skin temperature (TEMP). The sampling rates of BVP, HR, EDA, and TEMP signals were 64 Hz, 1 Hz, 4 Hz, and 4 Hz, respectively.
User events	Moody logs	Searching and listening records and timestamped user events (e.g., play/pause music, skip a track).
Emotion ratings	Pop-up survey	Arousal and valence rating [−10, 10] reported at the end of each music piece.

Participants’ emotional status reported after listening to each music piece was also aligned with their listening records. The arousal and valence ratings were subsequently grouped into three categories (i.e., positive, negative, neutral) and taken as the ground truth labels for the classification experiment. Specifically, ratings higher than zero were coded as positive, while those lower were negative, which resulted in 571 positive, 365 negative, and 20 neutral (i.e., 0) ratings in arousal and 647 positive, 283 negative, and 26 neutral (i.e., 0) ratings in valence. The imbalance of samples across categories may create challenges in experimentation for which resampling procedures could be adopted (see Section 5.1 for more details). The dataset with aligned music and physiological signals was named "HKU956" and deposited in our institutional data repository (The HKU956 dataset is available at: https://doi.org/10.25442/hku.21080821, accessed on 14 September 2022) for other researchers to conduct further analysis.

4. Feature Extraction

With physiological signal chunks and audio samples aligned with participants' listening records and emotion ratings, we further extracted a series of physiological and acoustic features to construct the feature sets for the MER tasks. In particular, physiological signals processing and feature extraction were based on the literature of physiological computing, bridging a knowledge gap in MIR.

4.1. Physiological Features

For controlling motion artifacts, filters were applied to the BVP and EDA signals prior to the physiological feature extraction. Specifically, an order-4 band-pass Butterworth filter was used to filter the BVP signals inside the passband of 1–8 Hz [42]. A Hanning moving average filter was applied to EDA signals with a window of 2.5 s [6]. As physiological signals vary across individuals, we further normalized the signals within each individual participant using z-score normalization [43]. We subsequently extracted a series of generic features from the preprocessed signals, including descriptive statistics of the physiological data (e.g., mean, standard deviation, min, max) as well as mean of the first/second differences of the time-series of BVP, EDA, HR, and TEMP signals on their original and normalized scales. In addition to these generic features, some physiological signal specific features were extracted as well.

Electrodermal activity (EDA). The EDA signals comprise two components: the slowly changing tonic component, which reflects a person's general skin conductance level (SCL), and the quickly changing phasic component, which is often elicited by external stimuli and is also referred to as the skin conductance response (SCR) [43]. We decomposed the EDA signal into its tonic and phasic component through Greco et al.'s cvxEDA approach [44] and extracted the following features from the two separate components (Table 3). For the tonic component, we computed the mean and standard deviation of the tonic activity level. For the phasic component, we computed the number and the rate of the phasic activity (i.e., the number of SCR per second), the mean and the standard deviation of the amplitude of the SCR peaks, and the mean and standard deviation of the SCR rise time and recovery time as in [45].

Table 3. EDA-specific features.

Category	Code	Description
Tonic	SCL	Mean and standard deviation of the tonic skin conductance level.
Phasic	SCR_num	The total number of detected SCRs.
	SCR_rate	The number of SCR per second.
	SCR_peak	Mean and standard deviation of the amplitude of the SCR peaks.
	SCR_rise_time	Mean and standard deviation of the time intervals between SCR onset and SCR peak.
	SCR_resp_time	Mean and standard deviation of the time intervals between SCR peak and recovery point (i.e., 50% of SCR amplitude).

Interbeat interval (IBI). To obtain the NN intervals (normal interbeat intervals) from the raw IBI signals, we removed abnormal intervals and ectopic beats based on Malik rules [46] and used linear interpolation to replace the outliers. A series of time- and frequency-domain analysis of NN intervals (NNI) were subsequently performed to measure participants' heart rate variability [47]. Specifically, for time domain features, apart from some descriptive statistics of the NN intervals (e.g., mean, standard deviation, median, range), we computed the SDSD, RMSSD, NN50, pNN50, NN20, pNN20, CVSD, CVNN from the NNI signals as well. Moreover, we also extracted a set of HRV features in the frequency domain, including total power, VLF, LF, HF, LF/HF, LFNU, and HFNU. Details on the computation of HRV features are described in Table 4.

Table 4. Heart rate variability features.

Category	Code	Description
Time domain	SDSD	Standard deviation of successive differences between adjacent NN intervals.
	RMSSD	The root mean square of successive differences between adjacent NN intervals.
	NN50, NN20	Number of pairs of adjacent NN intervals differing by more than 50 or 20 milliseconds.
	pNN50, pNN20	Percentage of differences between adjacent NN intervals that exceed 50 or 20 milliseconds.
	CVSD	The coefficient of variation of successive differences, i.e., RMSSD divided by the mean of the NN intervals.
	CVNN	The coefficient of variation, i.e., the ratio of SDNN divided by the mean of the NN intervals.
Frequency domain	Total power	Total power spectral density.
	VLF	Power in the very-low-frequency band (i.e., 0.003–0.04 Hz).
	LF	Power in the low-frequency band (i.e., 0.04–0.15 Hz).
	HF	Power in the high-frequency band (i.e., 0.15–0.4 Hz).
	LF/HF	The ratio of the power in the low-frequency band to that in the high-frequency band.
	LFNU	Low-frequency power in normalized units.
	HFNU	High-frequency power in normalized units.

4.2. Acoustic Features

Five categories of acoustic features corresponding to five major music characteristics were extracted, including loudness, rhythm, timbre, pitch, and harmony. The features and their dimensionalities are summarized in Table 5 and explained as follows.

Table 5. Extracted Acoustic Features.

Category	Feature	Dimensions
Loudness	Root-mean-square energy	4
Rhythm	Tempo	1
	Rhythm strength	1
	Global onset autocorrelation	2
	Average onset frequency	1
Timbre	Mel-frequency cepstral coefficient	38
	Δ MFCC, $\Delta\Delta$ MFCC	76
	Spectrum characteristics	22
	Zero crossing rate	2
Harmony	Tonal centroid (tonnetz)	12
Pitch	STFT chromagram	24
	Constant-Q chromagram	24
	CENS chromagram	24
Total		231

Loudness and dynamics generally represent the intensity of a sound and the variation of it. In this study, the root-mean-square energy (RMSE) was adopted as the loudness feature and was computed from the Short-Time Fourier Transform (STFT) spectrogram. In addition, we also computed the logarithmic compression version of RMSE to simulate human perception [48].

Rhythm, which depicts the pulse of music and the pattern of arrangement of musical notes, was characterized by the estimated tempo (i.e., beats per minute), rhythm strength (i.e., the average of the onset strength envelope), global onset autocorrelation (i.e., autocorrelation computed on the onset strength envelope), and the average onset frequency (i.e., the number of notes per second) [49].

Timbre, another music trait widely exploited in content-based MER, refers to the texture of a musical sound. Spectrum characteristics, zero crossing rate of the time domain audio signals, and Mel-frequency cepstral coefficient (MFCC) related features were used for this music trait dimension. Specifically, the spectrum characteristics involved in this study were roll-off frequency, flatness, spectral centroid, spectral contrast, and spectral bandwidth. Furthermore, we took the first 20 MFCCs and excluded the direct current (DC) term. The first and second order derivatives of MFCCs (i.e., Δ MFCC and $\Delta\Delta$ MFCC) were computed as well.

Pitch and harmony. Three versions of chromagrams were computed to represent the energy distribution of the 12 pitch classes (from C, C#, . . . to B), including a chromagram computed from the STFT spectrogram, a chromagram computed from Constant-Q Transform (CQT) spectrogram, and Chroma Energy Normalized Statistics (CENS). In the chromagram, each pitch class (also known as a chroma) embraces a set of pitch bands separated by octaves (e.g., the chroma C consists of {C1, C2, C3, C4, . . . }). The CQT chromagram further mirrors the human auditory sensation through logarithmic transformation, while the CENS chromagram smooths the local deviations and hence shows better robustness to dynamical and timbral variations [49]. Moreover, this study also included the tonal centroid (tonnetz) features following [50], which indicates the harmonic change of music. Specifically, based on the harmonic network, collections of pitch classes were projected to a six-dimensional space as tonal centroid points, which can be visualized on the circles of major thirds, minor thirds, and fifths. The tonnetz features are particularly useful for chord recognition [50].

The aforementioned acoustic features were all extracted using a specialized Python library for music and audio signal analysis, Librosa [51], with the default parameter settings (e.g., sample rate: 22050, hop length: 512). Unless otherwise specified, the aggregate statistics (i.e., mean, standard deviation) of the frame-wise feature vectors were used to form the clip-level feature representation.

5. Supervised Classification

Based on the aforementioned feature sets, we performed a set of classification experiments to investigate the potential contributions of physiological signals and music acoustic features as well as the effectiveness of group-wise modeling. Details on the classification experiments and the validation process are described in this section.

5.1. Classification Experiments

We built support vector machine (SVM) classifiers with the radial basis function (RBF) kernel for arousal and valence recognition, since it has demonstrated its effectiveness for MER tasks in previous research [52] including those based on physiological signals [37,53]. Given that only a small proportion of samples in our dataset were labeled as neutral arousal (2.1%) or neutral valence (2.7%), all classification experiments were thus implemented as a binomial classification problem (positive valence/arousal versus negative valence/arousal), in which we excluded the samples with neutral labels in this study, resulting in 936 samples for arousal prediction and 930 samples for valence classification.

In response to RQ1, we trained SVM models with three feature sets (i.e., audio-only, physiology-only, and audio + physiology) and compared the classification performance with Wilcox's robust paired-samples *t*-tests [54] and the Benjamini–Hochberg procedure [55] for controlling potential Type I errors introduced by multiple comparisons. Considering the imbalanced sample distribution across the positive and negative arousal/valence categories, apart from the experiments on the original imbalanced dataset, we also built prediction models with resampled balanced training data. Specifically, for each set of training folds, we undersampled the majority class via random undersampling and oversampled the minority class using the SVM borderline synthetic minority oversampling technique (SVM-SMOTE), which extrapolated minority instances along the SVM decision boundary [56]. In response to RQ2, we trained SVM classifiers on subgroups of users to investigate the role of users' personality in the customized modeling of music-induced emotion.

5.2. Validation

A set of 10-fold cross validations were performed for the validation of classifier performance. Particularly, for each classification experiment (i.e., classifiers trained on each feature set with either imbalanced or resampled data), the 10-fold cross validation was repeated 10 times with different splits of the dataset, since this process is deemed to be more robust especially when there is a resampling procedure involved in the classification pipeline [55]. As the testing folds were imbalanced, the evaluation metrics used were macro F1-score, which gives equal weight to each category, and AUC (area under ROC curve) [57].

6. Results and Discussion

6.1. Comparison of Different Modalities of Features

Table 6 presents classification performances on three modalities of features (i.e., audio-only, physiology-only, and audio + physiology) for arousal and valence prediction.

Table 6. Comparison of classification performance on different modalities.

Feature Set	Imbalanced Data				Resampled Data			
	Arousal		Valence		Arousal		Valence	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC
Audio-only	71.38%	81.31%	54.74%	63.86%	72.93%	81.41%	58.09%	63.72%
Physiology-only	47.64%	58.60%	54.14%	57.44%	55.79%	58.85%	57.07%	59.62%
Combined	**71.67%**	**81.42%**	**58.92%**	**70.26%**	**73.05%**	**81.46%**	**62.59%**	**68.96%**
Aud. vs. phy.	− ***	− ***	−	− ***	− ***	− ***	−	− ***
Aud. vs. comb.	+	+	+ ***	+ ***	+	+	+ ***	+ ***
Phy. vs. comb.	+ ***	+ ***	+ ***	+ ***	+ ***	+ ***	+ ***	+ ***

Note: The best performance in each column is highlighted with bold font. The Benjamini–Hochberg procedure was applied to control Type I error when comparing classification performance (*** $p < 0.001$). The "+"/"−" signs indicate the latter feature set performed better/worse than the former one. Aud. = Audio-only, Phy. = Physiology-only, Comb. = Combined.

The classification performance (in macro F1-score and AUC) showed a similar pattern across experiments with imbalanced or resampled training data. The combined multimodal feature set achieved the best performance in each experiment condition, while the improvement for valence prediction was more significant than that for arousal prediction.

Specifically, for *valence* classification, the combined feature set consistently outperformed the single-modality feature sets (i.e., audio-only, physiology-only) with statistical significance (at $p < 0.001$). In comparing the single-modality feature sets, the acoustic features showed slightly better performance than the physiological features. The relative strength of acoustic features was significant in AUC but not in the macro F1-score. For *arousal* prediction, though the combined feature set was found to be the best-performing, it did not significantly outperform the acoustic features (the best-performing single-modality feature set). Finally, the audio-only feature set showed a statistically significant advantage over the physiology-only feature set.

To sum up, with regard to RQ1, our results suggest that: for valence prediction, features extracted from peripheral physiological signals from listeners might be mediocre by themselves, but, when combined with acoustic features extracted from music content, the multimodal feature set showed significant improvement compared to feature sets of single modalities. For arousal prediction, the performances of physiological signals were in line with those on valence prediction, but little contribution was observed from the physiological features in the performance of the combined feature set. In other words, the acoustic modality alone was comparable with the multimodal approach.

Results of acoustic features shown in Table 6 are also in line with the findings of the previous research that acoustic features extracted from the music content are effective for arousal classification but less successful for valence prediction [25,49]. Associations between the dynamical and rhythmic characteristics of music and the induced arousal of listeners have been evidenced by studies in music psychology literature, while listeners' valence was also found to be influenced by some nonmusical factors as well, such as individuals' aesthetic judgement [58]. Hence, the acoustic features might be more comprehensive in modeling listeners' arousal than modeling listeners' valence. On the other hand, for the more challenging valence classification task, neither acoustic nor physiological feature set alone performed well, but the combined feature set significantly improved the classification performance over either single modality, indicating that the acoustic and physiological feature sets compensated for one another in valence classification. In contrast to previous studies that used only acoustic features, these findings reveal the potential values of physiological signals in tasks related to music-induced emotion. Future research could extend this work to further explore the applicability and discuss the balance between the cost and the benefits of incorporating physiological signals into an MIR system.

6.2. Group-Wise Classification Performance

To answer RQ2, whether users' personality plays a role in the predictive modeling of music-induced emotion, we trained SVM classifiers on the datasets partitioned by participants' personality (see Table 1), with the procedure described in Section 3.2. Particularly, for the group-wise classifications, we only reported the performances in AUC for experiments with resampled training data, as (1) macro F1-score and AUC reflected similar patterns on the whole dataset; (2) the results of experiments with imbalanced or resampled training data revealed a similar pattern on the whole dataset; and (3) the models trained on resampled data showed more stable performances.

Table 7 presents the classification performances (in AUC) of different feature modalities within each user personality group. Interestingly, compared to the generic model built on the whole (resampled) dataset, the group-wise models improved classification performances for some of the participant groups (indicated by "+" signs next to the performances in Table 7) but decreased prediction performances for the others (indicated by "−" signs in Table 7). For example, arousal classification within the extraversion (low) group and valence classification within the openness (high) group performed better than those in the whole dataset (Table 6), while arousal classification within the extraversion (high) group and valence classification within the conscientiousness (low) group performed worse. These observations imply that there might be some moderating effects of personality in individuals' emotional response to music. In other words, listeners with certain personality traits (e.g., less extraverted) may be more sensitive to acoustic features of music explored in this study and may have higher level physiological responses to music. This result is consistent with the results of [59] showing that extraversion is positively related to certain responses to music (e.g., happiness and sadness). This is also related to previous findings in psychology that introverted students tended to be more affected by background music due to relatively higher cortical arousal than extraverted students [60]. Group-wise classification might be helpful for modeling the effects of personality, which could be further verified in future investigations, preferably with larger samples.

Table 7. Classification performance (in AUC) within each participant group.

Personality Trait	Arousal				Valence			
	Size	Acoustics	Physiology	Combined	Size	Acoustics	Physiology	Combined
O (Low)	515	80.59% −	55.01% − ***	**80.67%** −	512	**62.92%** −	47.86% − ***	61.47% − ***
O (High)	421	**80.76%** −	58.06% −	80.43% −	418	63.61% −	63.87% + ***	**72.29%** + ***
C (Low)	558	**83.43%** + **	57.10% −	83.30% + ***	555	59.68% − ***	53.68% − ***	**60.09%** − ***
C (High)	378	76.76% − ***	52.53% − ***	**77.37%** − **	375	60.11% − **	62.31% + *	**66.94%** −
E (Low)	394	**82.22%** +	59.60% +	81.64% + ***	386	**55.48%** − ***	52.02% − ***	55.03% − ***
E (High)	542	79.09% − **	56.38% − *	**79.35%** − **	544	60.91% − **	61.43% +	**68.39%** −
A (Low)	760	**80.57%** −	56.01% − **	80.54% −	754	65.01% +	59.84% +	**69.71%** +
A (High)	176	82.46% +	61.88% +	**82.49%** +	176	38.95% − ***	**49.08%** − ***	44.48% − ***
E-S (Low)	562	80.12% −	62.94% + ***	**80.20%** −	556	60.27% − ***	57.45% −	**67.12%** −
E-S (High)	374	**81.84%** +	50.01% − ***	81.82% +	374	59.52% − **	61.93% +	**63.77%** − ***

Note: The best performances of arousal and valence prediction within each participant group are highlighted with bold font. Note that "size" denotes the sample size in terms of music pieces. The "+"/"−" signs indicate the performance was better/worse than that on the whole dataset (resampled) presented in Table 6. The Benjamini–Hochberg procedure was applied to control Type I error when comparing classification performance (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). O = Openness, C = Conscientiousness, E = Extraversion, A = Agreeableness, E-S = Emotional Stability.

Similar to the results on the whole dataset, we also compared group-wise classification performances of different modalities, with results presented in Table 8. In group-wise *arousal* classification experiments, the audio-only feature set consistently outperformed the physiology-only one, while the multimodal feature set did not significantly improve the performances of audio-only features in most cases (as indicated in the "Aud. vs. Comb." column in Table 8). This is consistent with our finding on the whole dataset that the acoustic features were effective for arousal classification and the music modality alone was comparable with the multimodal approach (cf. Section 6.1). On the other hand, in group-wise *valence* classification experiments, the combined feature set (i.e., audio + physiology) outperformed the audio-only feature set in all but two groups (i.e., openness to new experiences (low), extraversion (low)), and the improvement was statistically significant in the majority of cases (at $p < 0.001$, as indicated in the "Aud. vs. Comb." column in Table 8). In particular, for the two user groups where the audio-only feature set achieved better performance in valence prediction, the audio-only feature set significantly outperformed the other two feature sets in the openness (low) group (at $p < 0.001$), while, in the extraversion (low) group, the acoustic features outperformed the physiological features (at $p < 0.05$), but the performance difference between the audio-only and the multimodal feature set was not statistically significant. These observations may be related to the fact that individuals' music perception and music preference were found to be closely related to an array of personality trait dimensions, especially openness to new experiences [3,60], which is sensitive to art and beauty [60]. In a more general sense, this result verifies previous findings that people's responses to music result from multiple factors, some of which are relatively changeable and short-term such as physiological signals, while others are more stable and long-term such as personality traits [61], which further implies the importance of personalized modeling of music-induced emotion.

With respect to the classification experiments using feature sets of single modalities (i.e., audio-only vs. physiology-only), the audio-only feature set achieved better performance with statistical significance (at $p < 0.001$) for arousal prediction in most cases, while neither modality showed apparent advantage for valence prediction.

Table 8. Comparison of classification performances on different modalities in group-wise modeling.

Personality	Arousal			Valence		
	Aud. vs. Phy.	**Aud. vs. Comb.**	**Phy. vs. Comb.**	**Aud. vs. Phy.**	**Aud. vs. Comb.**	**Phy. vs. Comb.**
Openness (L)	$-^{***}$	$+$	$+^{***}$	$-^{***}$	$-^{***}$	$+^{***}$
Openness (H)	$-^{***}$	$-^{*}$	$+^{***}$	$+$	$+^{***}$	$+^{***}$
Conscientiousness (L)	$-^{***}$	$-$	$+^{***}$	$-^{***}$	$+$	$+^{***}$
Conscientiousness (H)	$-^{***}$	$+^{**}$	$+^{***}$	$+$	$+^{***}$	$+^{***}$
Extraversion (L)	$-^{***}$	$-^{***}$	$+^{***}$	$-^{*}$	$-$	$+^{**}$
Extraversion (H)	$-^{***}$	$+$	$+^{***}$	$+$	$+^{***}$	$+^{***}$
Agreeableness (L)	$-^{***}$	$-$	$+^{***}$	$-^{***}$	$+^{***}$	$+^{***}$
Agreeableness (H)	$-^{***}$	$+$	$+^{***}$	$+^{***}$	$+^{***}$	$-^{*}$
Emotional stability (L)	$-^{***}$	$+$	$+^{***}$	$-^{*}$	$+^{***}$	$+^{***}$
Emotional stability (H)	$-^{***}$	$-$	$+^{***}$	$+$	$+^{***}$	$+$

Notes: The Benjamini–Hochberg procedure was applied to control Type I error when comparing classification performance (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). The "+"/"−" signs indicate the latter feature set performed better/worse than the former one. Aud. = Audio-only, Phy. = Physiology-only, Comb. = Combined.

7. Conclusions, Limitations, and Future Works

Aiming to explore the potentials for incorporating physiological signals for detecting music-induced emotion and the role of users' personality in group-wise modeling, this study constructed a thorough dataset with aligned music stimuli audio and listeners' physiological signals, ratings on emotion, and personalities. To bridge the gap between physiological computing and MIR, this study leveraged techniques in physiological signal processing to extract physiological signals that are potentially interpretable. Our findings on multimodal music emotion classification make it evident that users' physiological responses to music is an informative modality for the MER task on music-induced emotion and made more significant contributions to valence than arousal classification. The physiology modality was found to be complementary to the acoustic modality for valence prediction. However, for arousal classification, the acoustic modality alone was comparable with the multimodal feature set. Additionally, the results of our user-group-wise classification experiment imply that both music acoustic and listeners' physiological signals might be more effective in predicting music-induced emotions for users of certain personality traits than other users. These observations imply moderating effects of personality in individuals' emotional response to music.

Through a theoretical lens, our findings call for a more holistic perspective when studying music-induced emotion. Despite the proved effect of music characteristics on listeners' arousal levels, its effect on valence response can be largely individual-dependent. Beyond psychoacoustic analysis, there are multiple factors contributing to inducing emotion in music listeners, some of which are relatively changeable and short-term (e.g., physiological signals), while others are more stable and long-term (e.g., personality traits). Practically, through demonstrating the potential of physiological sensing techniques in emotion-aware MIR, this study opens up a number of possibilities in the future for emotion-aware or context-based music retrieval systems, such as recommending music based on physiological metrics of users and maintaining emotion-sensitive playlists which can be adjusted in real time in accordance with users' changing physiological response to music. In addition, findings of this study also highlight the importance of multimodal and personalized modeling that exploits not only content-based MER but also various forms of user-dependent information sources (e.g., personality, physiological signals). Findings of this line of research may also prompt academic and industrial advancement in closely related areas such as human physiological response, human-computer interaction, music perception, and music therapy.

This study has several limitations. Although the HKU956 dataset is of a decent size, samples with neutral labels were finally discarded, which resulted in smaller samples, especially in the group-wise experiments. Moreover, our dataset also faced the issue of being dominated by the positive class, especially for the valence dimension. This data unbalance toward positive valence might be attributable to the experiment design where participants searched and selected music to listen to rather than listening to preselected music. While this experiment design addresses the MIR contexts, it would result in more music favorable to the listeners being included in the dataset. Future work could explore personalized models if larger datasets could be constructed with more listening history and corresponding physiological signals from each user. In the experiment setup, album covers and song metadata were displayed on screen, albeit in small sizes, when music was played. This setup conforms to the norm of online music services but does introduce potential influence from visual information on listeners' perception of the song. Future studies may avoid this by displaying a blank screen when music is played, trading off the ecological authenticity of the user interface.

Given that different physiological signals may reinforce each other, future research is much needed to investigate the contribution each physiological signal makes to the outcome. More work is also suggested to exploit other personal and situational factors beyond personality that may play a role in ones' affective perception of music, such as music training background and/or tasks at hand. Last but not least, it is noteworthy that, as physiological signals and audio signals are both time series data, a multimodal time series approach could thus be exploited for many interesting research topics, such as detecting variation of music-induced emotion within the course of a music piece.

Author Contributions: Conceptualization, X.H.; Data curation, X.H. and R.L.; Formal analysis, F.L.; Funding acquisition, X.H.; Investigation, X.H. and R.L.; Methodology, X.H.; Project administration, X.H.; Software, X.H. and R.L.; Supervision, X.H.; Validation, X.H. and F.L.; Writing—original draft, X.H. and F.L.; Writing—review and editing, X.H., F.L. and R.L. All authors have read and agreed to the published version of the manuscript.

Funding: This study is supported by National Natural Science Foundation of China (No. 61703357) and the Research Grants Council of the Hong Kong S.A.R., China (No. HKU 17607018).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Human Research Ethics Committee (HREC) of The University of Hong Kong (protocol code EA1802092, approved on 29 March 2018).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study ("HKU956") are openly available in HKU Data Repository at https://doi.org/10.25442/hku.21080821.

Acknowledgments: The authors thank the reviewers for the detailed comments which are very valuable for improving the manuscript. The participants of the user experiment are also appreciated for their contribution to this study.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Abdul, A.; Chen, J.; Liao, H.Y.; Chang, S.H. An emotion-aware personalized music recommendation system using a convolutional neural networks approach. *Appl. Sci.* **2018**, *8*, 1103. [CrossRef]
2. Dunker, P.; Nowak, S.; Begau, A.; Lanz, C. Content-based mood classification for photos and music: A generic multi-modal classification framework and evaluation approach. In Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, Vancouver, BC, Canada, 30–31 October 2008; pp. 97–104.
3. Rentfrow, P.J.; Gosling, S.D. The Do Re Mi's of everyday life: The structure and personality correlates of music preferences. *J. Personal. Soc. Psychol.* **2003**, *84*, 1236–1256. [CrossRef] [PubMed]
4. North, A.C.; Hargreaves, D.J. Liking, Arousal potential, and the emotions expressed by music. *Scand. J. Psychol.* **1997**, *38*, 45–53. [CrossRef]
5. Kallinen, K.; Ravaja, N. Emotion perceived and emotion felt: Same and different. *Musicae Sci.* **2006**, *10*, 191–213. [CrossRef]
6. Picard, R.W.; Vyzas, E.; Healey, J. Toward machine emotional intelligence: Analysis of affective physiological State. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 1175–1191. [CrossRef]
7. Sivathasan, S.; Philibert-Lignières, G.; Quintin, E.M. Individual differences in autism traits, personality, and emotional responsiveness to music in the general population. *Musicae Sci.* **2021**, *26*, 1029864920988160. [CrossRef]

8. Agres, K.R.; Schaefer, R.S.; Volk, A.; van Hooren, S.; Holzapfel, A.; Dalla Bella, S.; Müller, M.; De Witte, M.; Herremans, D.; Ramirez Melendez, R.; et al. Music, computing, and health: A roadmap for the current and future roles of music technology for health care and well-being. *Music Sci.* **2021**, *4*, 2059204321997709. [CrossRef]
9. Hu, X.; Chen, J.; Wang, Y. University students' use of music for learning and well-being: A qualitative study and design implications. *Inf. Process. Manag.* **2021**, *58*, 102409. [CrossRef]
10. Hu, X.; Downie, J.S.; Laurier, C.; Bay, M.; Ehmann, A.F. The 2007 MIREX audio mood classification task: Lessons learned. In Proceedings of the 9th International Conference on Music Information Retrieval, Philadelphia, PA, USA, 14–18 September 2008; pp. 462–467.
11. Aljanaki, A.; Yang, Y.-H.; Soleymani, M. Emotion in music task at MediaEval 2015. In Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, 14–15 September 2015.
12. Gómez-Cañón, J.S.; Cano, E.; Eerola, T.; Herrera, P.; Hu, X.; Yang, Y.H.; Gómez, E. Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications. *IEEE Signal Process. Mag.* **2021**, *38*, 106–114. [CrossRef]
13. Juslin, P.N.; Laukka, P. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *J. New Music Res.* **2004**, *33*, 217–238. [CrossRef]
14. Juslin, P.N. Communicating emotion in music performance: A review and a theoretical framework. In *Music and Emotion: Theory and Research*; Juslin, P.N., Sloboda, J.A., Eds.; Oxford University Press: New York, NY, USA, 2001; pp. 309–337.
15. Wang, X.; Wang, L.; Xie, L. Comparison and analysis of acoustic features of western and Chinese classical music emotion recognition based on VA model. *Appl. Sci.* **2022**, *12*, 5787. [CrossRef]
16. Hu, X.; Choi, K.; Downie, J.S. A framework for evaluating multimodal music mood classification. *J. Assoc. Inf. Sci. Technol.* **2017**, *68*, 273–285. [CrossRef]
17. Aljanaki, A.; Wiering, F.; Veltkamp, R. Computational modeling of induced emotion using GEMS. In Proceedings of the 15th Conference of the International Society for Music Information Retrieval (ISMIR 2014), Taipei, Taiwan, 27–31 October 2014; pp. 373–378.
18. Jakubik, J.; Kwaśnicka, H. Sparse coding methods for music induced emotion recognition. In Proceedings of the 2016 Federated Conference on Computer Science and Information Systems (FedCSIS), Gdansk, Poland, 11–14 September 2016; pp. 53–60.
19. Chang, W.; Li, J.; Lin, Y.; Lee, C. A Genre-Affect Relationship Network with Task-Specific Uncertainty Weighting for Recognizing Induced Emotion in Music. In Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, USA, 23–27 July 2018; pp. 1–6.
20. Avramidis, K.; Garoufis, C.; Zlatintsi, A.; Maragos, P. Enhancing affective representations of music-induced EEG through multimodal supervision and latent domain adaptation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 4588–4592. [CrossRef]
21. Daly, I.; Williams, D.; Hallowell, J.; Hwang, F.; Kirke, A.; Malik, A.; Weaver, J.; Miranda, E.; Nasuto, S.J. Music-induced emotions can be predicted from a combination of brain activity and acoustic features. *Brain Cogn.* **2015**, *101*, 1–11. [CrossRef]
22. Bhatti, A.M.; Majid, M.; Anwar, S.M.; Khan, B. Human emotion recognition and analysis in response to audio music using brain signals. *Comput. Human Behavior* **2016**, *65*, 267–275. [CrossRef]
23. Daly, I.; Nicolaou, N.; Williams, D.; Hwang, F.; Kirke, A.; Miranda, E.; Nasuto, S.J. Neural and physiological data from participants listening to affective music. *Sci. Data* **2020**, *7*, 177. [CrossRef] [PubMed]
24. Hsu, Y.; Wang, J.; Chiang, W.; Hung, C. Automatic ECG-based emotion recognition in music listening. *IEEE Trans. Affect. Comput.* **2017**, *11*, 85–99. [CrossRef]
25. Coutinho, E.; Cangelosi, A. Musical emotions: Predicting second-by-second subjective feelings of emotion from low-level psychoacoustic features and physiological measurements. *Emotion* **2011**, *11*, 921–937. [CrossRef]
26. Greer, T.; Ma, B.; Sachs, M.; Habibi, A.; Narayanan, S. A multimodal view into music's effect on human neural, physiological, and emotional experience. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 167–175.
27. Healey, J.; Picard, R.; Dabek, F. A new affect-perceiving interface and its application to personalized music selection. In Proceedings of the 1998 Workshop on Perceptual User Interfaces, San Fransisco, CA, USA, 5–6 November 1998; pp. 4–6.
28. Janssen, J.H.; van den Broek, E.L.; Westerink, J.H.D.M. Tune in to your emotions: A robust personalized affective music player. *User Modeling User-Adapt. Interact.* **2012**, *22*, 255–279. [CrossRef]
29. van der Zwaag, M.D.; Janssen, J.H.; Westerink, J.H.D.M. Directing physiology and mood through music: Validation of an affective music player. *IEEE Trans. Affect. Comput.* **2013**, *4*, 57–68. [CrossRef]
30. Chiu, M.-C.; Ko, L.-W. Develop a personalized intelligent music selection system based on heart rate variability and machine learning. *Multimed. Tools Appl.* **2017**, *76*, 15607–15639. [CrossRef]
31. Kim, H.G.; Lee, G.Y.; Kim, M.S. Dual-function integrated emotion-based music classification system using features from physiological signals. *IEEE Trans. Consum. Electron.* **2021**, *67*, 341–349. [CrossRef]
32. Hsu, J.-L.; Zhen, Y.-L.; Lin, T.-C.; Chiu, Y.-S. Affective content analysis of music emotion through EEG. *Multimed. Syst.* **2018**, *24*, 195–210. [CrossRef]
33. Hu, X.; Li, F.; Ng, T.D.J. On the relationships between music-induced emotion and physiological signals. In Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR '18), Paris, France, 23–27 September 2018; pp. 362–369.

34. Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I. DEAP: A database for emotion analysis; using physiological signals. *IEEE Trans. Affect. Comput.* **2011**, *3*, 18–31. [CrossRef]
35. Zhang, K.; Zhang, H.; Li, S.; Yang, C.; Sun, L. The PMEmo dataset for music emotion recognition. In Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, Yokohama, Japan, 11–14 June 2018; pp. 135–142.
36. Sangnark, S.; Autthasan, P.; Ponglertnapakorn, P.; Chalekarn, P.; Sudhawiyangkul, T.; Trakulruangroj, M.; Songsermsawad, S.; Assabumrungrat, R.; Amplod, S.; Ounjai, K.; et al. Revealing preference in popular music through familiarity and brain response. *IEEE Sens. J.* **2021**, *21*, 14931–14940. [CrossRef]
37. Chaturvedi, V.; Kaur, A.B.; Varshney, V.; Garg, A.; Chhabra, G.S.; Kumar, M. Music mood and human emotion recognition based on physiological signals: A systematic review. *Multimed. Syst.* **2022**, *28*, 21–44. [CrossRef]
38. Hu, X.; Lee, J.; Bainbridge, D.; Choi, K.; Organisciak, P.; Downie, J.S. The MIREX Grand Challenge: A framework of holistic user experience evaluation in music information retrieval. *J. Assoc. Inf. Sci. Technol.* **2017**, *68*, 97–112. [CrossRef]
39. Russell, J.A.; Weiss, A.; Mendelsohn, G.A. Affect grid: A single-item scale of pleasure and arousal. *J. Personal. Soc. Psychol.* **1989**, *57*, 493. [CrossRef]
40. Gosling, S.D.; Rentfrow, P.J.; Swann, W.B. A very brief measure of the Big-Five personality domains. *J. Res. Personal.* **2003**, *37*, 504–528. [CrossRef]
41. Navarro-Alamán, J.; Lacuesta, R.; García-Magariño, I.; Lloret, J. EmotIoT: An IoT system to improve users' wellbeing. *Appl. Sci.* **2022**, *12*, 5804. [CrossRef]
42. Martinho, M.; Fred, A.; Silva, H. Towards continuoususer recognition by exploring physiological multimodality: An electrocardiogram (ECG) and blood volume pulse (BVP) approach. In Proceedings of the 2018 International Symposium in Sensing and Instrumentation in IoT Era (ISSI), Shanghai, China, 6–7 September 2018; pp. 1–6.
43. Gashi, S.; Lascio, E.D.; Santini, S. Using unobtrusive wearable sensors to measure the physiological synchrony between presenters and audience members. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2019**, *3*, 13. [CrossRef]
44. Greco, A.; Valenza, G.; Lanata, A.; Scilingo, E.P.; Citi, L. cvxEDA: A convex optimization approach to electrodermal activity processing. *IEEE Trans. Biomed. Eng.* **2015**, *63*, 797–804. [CrossRef]
45. Zhang, L.; Wade, J.; Bian, D.; Fan, J.; Swanson, A.; Weitlauf, A.; Warren, Z.; Sarkar, N. Cognitive load measurement in a virtual reality-based driving system for Autism intervention. *IEEE Trans. Affect. Comput.* **2017**, *8*, 176–189. [CrossRef] [PubMed]
46. Kamath, M.V.; Fallen, E.L. Correction of the heart rate variability signal for ectopics and missing beats. In *Heart Rate Variability*; Malik, M., Camm, A.J., Eds.; Futura Publishing Co., Inc.: Armonk, NY, USA, 1995; pp. 75–85.
47. Shaffer, F.; Ginsberg, J.P. An overview of heart rate variability metrics and norms. *Front. Public Health* **2017**, *5*, 258. [CrossRef] [PubMed]
48. Müller, M. Dynamics, intensity, and loudness. In *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*; Müller, M., Ed.; Springer International Publishing: Cham, Switzerland, 2015; pp. 24–26.
49. Hu, X.; Yang, Y. Crossdataset and cross-cultural music mood prediction: A case on western and Chinese pop songs. *IEEE Trans. Affect. Comput.* **2017**, *8*, 228–240. [CrossRef]
50. Harte, C.; Sandler, M.; Gasser, M. Detecting harmonic change in musical audio. In Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia, Santa Barbara, CA, USA, 27 October 2006; pp. 21–26.
51. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.; McVicar, M.; Battenberg, E.; Nieto, O. Librosa: Audio and music signal analysis in python. In Proceedings of the 14th Python in Science Conference, Austin, TX, USA, 6–12 July 2015; pp. 18–25.
52. Yang, X.; Dong, Y.; Li, J. Review of data features-based music emotion recognition methods. *Multimed. Syst.* **2018**, *24*, 365–389. [CrossRef]
53. Ayata, D.; Yaslan, Y.; Kamasak, M.E. Emotion based music recommendation system using wearable physiological sensors. *IEEE Trans. Consum. Electron.* **2018**, *64*, 196–203. [CrossRef]
54. Wilcox, R. *Introduction to Robust Estimation and Hypothesis Testing*, 3rd ed.; Academic Press: Boston, MA, USA, 2012; pp. 137–213.
55. Thissen, D.; Steinberg, L.; Kuang, D. Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *J. Educ. Behav. Stat.* **2002**, *27*, 77–83. [CrossRef]
56. Nguyen, H.M.; Cooper, E.W.; Kamei, K. Borderline over-sampling for imbalanced data classification. In Proceedings of the 5th International Workshop on Computational Intelligence & Applications, Hiroshima, Japan, 10–12 November 2009; pp. 24–29.
57. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [CrossRef]
58. Juslin, P.N. Emotional reactions to music. In *The Oxford Handbook of Music Psychology*, 2nd ed.; Hallam, S., Cross, I., Thaut, M., Eds.; Oxford University Press: Oxford, UK, 2016; pp. 197–213.
59. Vuoskoski, J.K.; Eerola, T. Measuring music-induced emotion: A comparison of emotion models, personality biases, and intensity of experiences. *Musicae Sci.* **2011**, *15*, 159–173. [CrossRef]
60. Küssner, M.B. Eysenck's theory of personality and the role of background music in cognitive task performance: A minireview of conflicting findings and a new perspective. *Front. Psychol.* **2017**, *8*, 1991. [CrossRef]
61. Schedl, M.; Flexer, A.; Urbano, J. The neglected user in music information retrieval research. *J. Intell. Inf. Syst.* **2013**, *41*, 523–539. [CrossRef]