Dual-Function Integrated Emotion-Based Music Classification System Using Features From Physiological Signals

Hyoung-Gook Kim, Gi Yong Lee, and Min-Soo Kim

***Abstract*—In this paper, we propose an emotion-based music classification system using features from physiological signals. The proposed system integrates two functions; the first uses physiological sensors to recognize the emotions of users listening to music, and the second classifies music according to the feelings evoked in the listeners, without using physiological sensors. Moreover, to directly predict the user's emotions from sensor data acquired through wearable physiological sensors, we developed and implemented a hierarchical inner attention-mechanism-based deep neural network. To relieve the discomfort of users wearing physiological sensors every time to receive content recommendations, the relation between emotion-specific features that are extracted from previously generated physiological signals, and musical features that are extracted from music is learned through a regression neural network. Based on these models, the proposed system classifies input music automatically according to users' emotional reactions without measuring human physiological signals. The experimental results not only demonstrate the accuracy of the proposed automatic music classification framework, but also provide a new perspective in which human experience-based characteristics related to emotion are applied to artificial-intelligence-based content classification.**

***Index Terms*—Deep neural network, emotion-based music classification, musical features, physiological signals, regression neural network.**

I. Introduction

MUSIC is an artistic medium that brings joy to people and expresses human thoughts and feelings through sound. Owing to the development of the Internet and the growth of the digital music market, the need for a music search and recommendation system [1] that can easily and quickly access various types of music from large music datasets has emerged. In conventional music search and recommendation systems, music is automatically classified based on genre [2], [3], associated emotion [4], [5], artist [6], lyrics [7], album [8], emotion displayed in videos [9], user profiling [10]–[12], content-based features [13]–[17], and users' social media interactions [18], and appropriate search and recommendation results are provided to users. Recent music recommendation models use variations of a hybrid system [19], [20] combining collaborative filtering [21], content-based filtering [22], context-based filtering [23]–[26], and metadata-based models [27] along with several other parameters. However, most music search and recommendation systems have been developed based on a system-centric rather than user-centric perspective; further, although emotions are an important factor in music selection, studies on emotions or expressions of music listeners remain insufficient.

Human emotions are an inherent part of our experience and depend on circumstantial causes and conditions in our lives. Manipulation of information related to users' emotions can be found in many areas of consumer electronics. For example, some cameras can detect facial expressions [28]–[30]; audio systems can automatically select types of music that are suitable for the user's mood [31] or an input image [32]; and monitoring devices can notify authorized persons of users' emotional changes [33]–[36]. Owing to the recent development of sensor technology, emotional characteristics have been expanded and applied as new features to consumer electronic devices to ensure more natural and intelligent interaction with humans [37]–[39]. Accordingly, affective computing [40] has garnered increasing attention as a research topic combining computer science with psychology and cognitive science.

Previous research on analyzing human emotions focused on speech recognition [41] and facial expression recognition [42]. In recent years, human emotion analysis by using physiological signals has emerged as an important research topic. Recognition of physiological signals [43]–[45] such as brainwaves, heart rate, electrocardiograms, skin conduction, respiration, body temperature, and pulse waves can provide insight into a person's emotional state, which is linked to their physical state. Hence, wearable devices with physiological sensors are increasingly being applied in fitness, gaming, disability, and medical industries in recent years. In addition, physiological signal analysis is being extended to potential applications such as power consumption management in air purifiers and interfaces in the consumer electronics field [37].

Deep learning techniques that can recognize users' emotions caused by music via physiological signals can maximize the satisfaction of listeners in various situations through application in music classification/recommendation systems [46].

Manuscript received May 28, 2021; revised August 16, 2021 and October 6, 2021; accepted October 8, 2021. Date of publication October 15, 2021; date of current version December 23, 2021. This work was supported in part by the Research Grant of Kwangwoon University in 2021 and in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2018R1D1A1B07041783. *(Corresponding author: Hyoung-Gook Kim.)*

The authors are with the Electronic Convergence Engineering Department, Kwangwoon University, Seoul 01897, Republic of Korea (e-mail: hkim@kw.ac.kr; agayong93@kw.ac.kr; ms1254@kw.ac.kr).

Digital Object Identifier 10.1109/TCE.2021.3120445

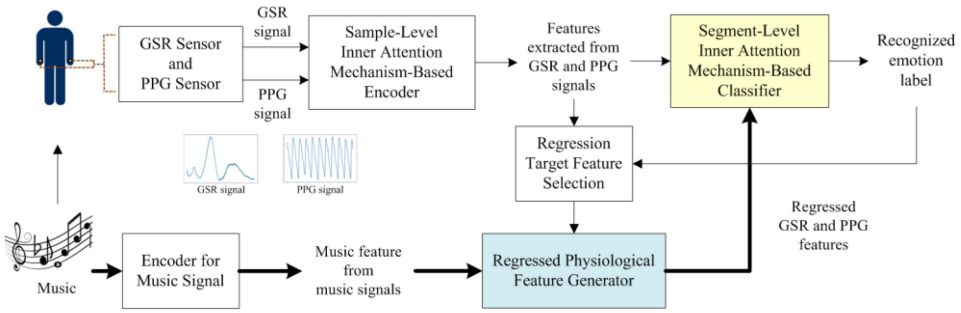

Fig. 1. Proposed system architecture for dual-function integrated emotion-based music classification system.

However, measuring physiological signals through wearable devices to achieve recommendation is significantly inconvenient.

In this paper, we propose an emotion-based music classification system that combines two functions using physiological features. The first function automatically classifies music by extracting the emotion patterns of users in response to a specific music source from physiological signals associated with that music. The other function automatically classifies the input music according to emotions without the use of sensors by learning the features of emotion patterns, which are extracted from physiological signals using machines designed to perform emotion recognition.

The contributions of this study are as follows.

1) An inner attention-mechanism-based deep neural network [47] is applied hierarchically to signals extracted from galvanic skin response (GSR) and photoplethysmography (PPG) sensors placed on participants when they listen to music to classify different emotions with high accuracy.

2) A multilayer perceptron (MLP) is used as a nonlinear regression function to ensure a strong correlation function between physiological features extracted during emotion recognition and musical features extracted from music data.

3) In the actual application, physiological features, which are mapped to the musical features of the input music based on MLP-based regression models, are automatically generated. Music is classified by applying these features to segment-level inner attention-mechanism-based deep neural networks.

II. System Architecture

Fig. 1 illustrates the process flow of the proposed emotion-based music classification system; it consists of Bluetooth bracelets with built-in PPG and GSR sensors and emotion recognition applications within a smartphone. In this study, we used GSR and PPG signals.

In the case where physiological sensors are applied, the signals obtained from the GSR and PPG sensors worn by music listeners are converted into the frequency domain. Then, they are used as input to a sample-level inner attention-mechanism-based convolutional recurrent neural network encoder to extract emotional features and send them to a smartphone. In the smartphone, the extracted features are applied to a bidirectional gated recurrent unit (BGRU) with segment-level inner attention mechanism models to classify the music by emotion.

In the case where physiological sensors are not applied, regression target features are selected among the previously generated GSR and PPG features, which are extracted from the sample-level inner attention-mechanism-based convolutional recurrent neural network encoder. The correlation model between the selected regression target features and musical features is then generated by the MLP regression training on the server and sent to the smartphone. During the application phase on the smartphone, musical features of the input music signal are entered into MLP-based regressors to automatically generate emotion features of PPG and GSR. Then, the fused GSR and PPG features are input into segment-level inner attention-mechanism-based bidirectional gated recurrent neural networks for emotion-based music classification.

A. Scalogram Sample-Level Inner Attention Mechanism-Based CBGRU Encoder

Sounds stimulate auditory organs (such as the ears), which transmit signals to the brain, and are reflected in changes to the heart and skin. When people hear sounds, they tend to listen attentively and intuitively to specific parts rather than all the available details. Based on this concept, we apply a convolutional bidirectional gated recurrent unit (CBGRU) network with scalogram sample-level inner attention mechanism to encode physiological sequences that respond to sound stimuli. The structure of the proposed encoder is shown in Fig. 2.

For preprocessing, after obtaining the GSR and PPG signals, each signal is divided into 0.5 s segments with 30% overlap. Then, each segment is converted into a two-dimensional scalogram [48] through wavelet transformation and fed into the CBGRU encoder. Time-series signals with a low sampling rate, such as physiological signals, have the advantage of providing better time-frequency localization in the wavelet domain than in the Fourier domain. As a result, wavelet-based graphic representations such as scalograms have been recently applied.

The CBGRU-based sample encoder, which is composed of a convolution layer and a BGRU layer, extracts features through learning the correlation between samples of scalograms from the GSR and PPG segments. The convolution operation is used to detect meaningful microstructure features, while maintaining the positional characteristics of the data. To achieve this, we use two convolutional layers here. The first convolutional

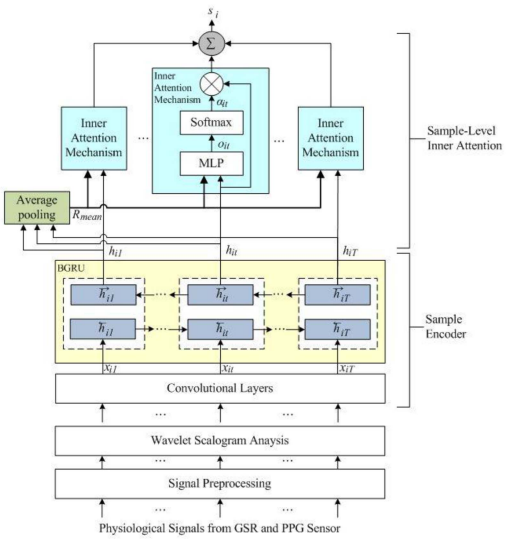

Fig. 2. Scalogram sample-level inner attention mechanism–based CBGRU (SSIA-CBGRU) encoder architecture.

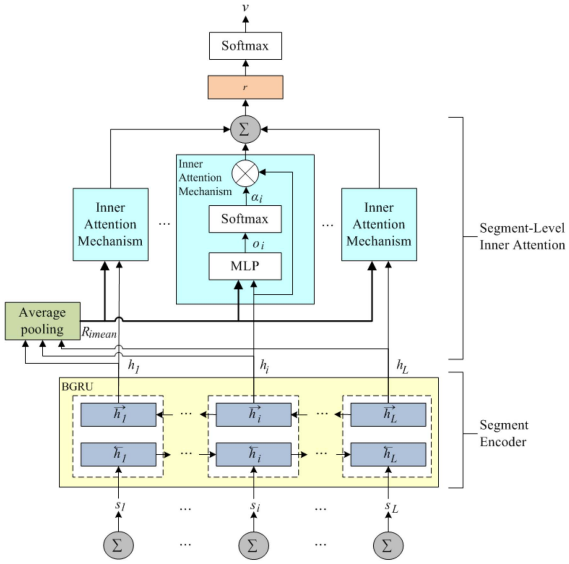

Fig. 3. BGRU architecture with segment-level inner attention mechanism model.

layer contains 16 kernel filters of size 3×3, and the second convolutional layer contains 32 kernel filters of size 3×3. Each of these convolutional layers was followed by three layers of batch normalization, rectified linear unit, and maximum pooling layers with kernel size of 2×2. Thereafter, the feature vectors extracted through the subsequent fully connected layer are used as the input to the BGRU; then, features that reflect the context interdependence of the feature vectors are extracted through learning. The BGRU extends unidirectional GRU networks by splitting each hidden layer into two separate layers. As shown in (1) and (2), one layer acts as a forward GRU and processes sequences x_{it} (representing the t^{th} sample in the i^{th} segment) in a temporal order (from x_{i1} to x_{iT}), and the other layer acts as a backward GRU in the opposite direction of the temporal order.

$$\vec{h}_{it} = G\vec{R}U_f(x_{it}), t \in [1, T], \tag{1}$$

$$\overleftarrow{h}_{it} = G\overleftarrow{R}U_b(x_{it}), t \in [T, 1], \tag{2}$$

This allows the BGRU encoder to extract powerful representations of the context of future and past time-series physiological features as follows.

$$\vec{h}_t = f(w_1 x_t + w_2 \vec{h}_{t-1}), \tag{3}$$

$$\overleftarrow{h}_t = f(w_3 x_t + w_5 \overleftarrow{h}_{t+1}), \tag{4}$$

$$h_t = g(w_4 \vec{h}_t + w_6 \overleftarrow{h}_t). \tag{5}$$

As a next step, a sample-level inner attention mechanism is applied to the output features from the CBGRU as follows.

$$o_{it} = \tanh(W_s h_{it} + W_m R_{mean} \otimes e_T), \tag{6}$$

$$\alpha_{it} = \frac{\exp(o_s^\intercal o_{it})}{\sum_t \exp(o_s^\intercal o_{it})}, \tag{7}$$

$$s_i = \sum_t h_{it} \alpha_{it}^\intercal \tag{8}$$

where W_s and W_m are the weight matrix of the single-layer MLP, and e_T is a vector of 1s equal to the number of T samples; s_i, h_{it}, R_{mean}, and α_{it} denote the segment context vector output from the inner attention mechanism, the output vector of the BGRU hidden layer, the average vector of all output vectors extracted from the BGRU hidden layer of each segment, and the inner annotation weight, respectively.

The sample-level inner attention mechanism computes more accurate and focused physiological features (O_{it}) that respond to sound stimuli by replacing the value computed through the average pooling of all output vectors of the BGRU in each segment instead of attention bias applied in the conventional method, and uses them as the annotations of the beneficial scalogram (α_{it}) to reflect them in the context vector representation (s_i) of the segment.

To calculate the normalized inner attention weight α_{it}, representing the importance of the segment sample x_{it}, first, as shown in (6), the output vector h_{it} of the BGRU hidden layer and average vector R_{mean} of h_{it} are used as input to a single-layer MLP, and the corresponding o_{it} is the output. Subsequently, the similarity between o_{it} and o_s is normalized through the softmax function (7) to obtain α_{it}. o_s is a context vector between samples, which is a parameter with an initial random value that is updated during model training. If the similarity is high, the inner attention weight is high. s_i is calculated using the sum of the weights of h_{it} of the BGRU hidden layer and the weight α_{it} of the scalogram sample annotation. In this case, the output of the hidden layer with a high inner attention weight is reflected in the segment vector s_i, i.e., the higher the corresponding inner attention weight, the larger the value is when the neural network learns.

B. Segment-Level Inner Attention-Mechanism-Based BGRU Encoder

Fig. 3 shows the architecture of the proposed segment-level inner attention-mechanism-based BGRU (SIA-BGRU)

classifier. The segment vector s_i is applied to this structure to predict the emotion of the living body sequence.

As shown in (9) and (10), the BGRU-based segment encoder takes the input from the segment vector s_i and calculates the annotation vector h_i by summarizing the forward $\vec{h}_i$and backward $\overleftarrow{h}_i$. The encoder process ensures that adjacent context information around segment i is incorporated into the annotation vector h_i in a similar manner as (5).

$$\vec{h}_i = G\vec{R}U_f(s_i), i \in [1, L], \tag{9}$$

$$\overleftarrow{h}_i = G\overleftarrow{R}U_b(s_i), i \in [L, 1], \tag{10}$$

where L is the number of segments in each sequence. The number of hidden units is also L.

Subsequently, the inner attention mechanism is again used to extract context information between segments, which improves classification accuracy of the physiological sequence as follows.

$$o_i = \tanh(W_e h_i + W_i R_{imean} \otimes e_L), \tag{11}$$

$$\alpha_i = \frac{\exp(o_e^T o_i)}{\sum_i \exp(o_e^T o_i)}, \tag{12}$$

$$r = \sum_i h_i \alpha_i^{\mathsf{T}} \tag{13}$$

where, e_L is a vector of 1s equal to the number of L segments. The averaged vectors R_{imean} of the annotation vector h_i of each segment and h_i are fed to the single-layer MLP, and the hidden representation o_i is calculated by (11). The normalized importance weight α_i is measured by applying the result of calculating the similarity between o_i and the segment-level context vector o_e to the softmax function, as shown in (12). Then, as in (13), a segment sequence vector r is obtained by the weighted sum of the segment-level inner attention weight α_i and each segment annotation vector h_i of the segment encoder, i.e., r is a high-level feature vector containing the overall information of the physiological sequences, which plays a more important role for emotion classification.

Lastly, for prediction of the probability distribution v, the segment sequence vector r is applied to the softmax layer as follows.

$$v = \text{softmax}(W_c r + b_c), \tag{14}$$

where W_c and b_c are parameters of the softmax layer.

The segment-level context vectors o_e, the basic weights W_e, W_i, W_c, and the bias vectors b_c, which are introduced to measure the importance of segments during the training process, are randomly initialized and fine-tuned. Thus, the inner attention output, r, is received as an input and the weight of whichever part is the most important is reflected in learning so that a more accurate result can be acquired. As for the optimization function, the Adam optimizer is used to minimize the cross-entropy error between the predicted label and the actual label for all segments. For the estimated losses, an optimization that adjusts the weight parameters and biases is performed. This process is repeated until the neural network reaches the desired accuracy or the highest possible accuracy.

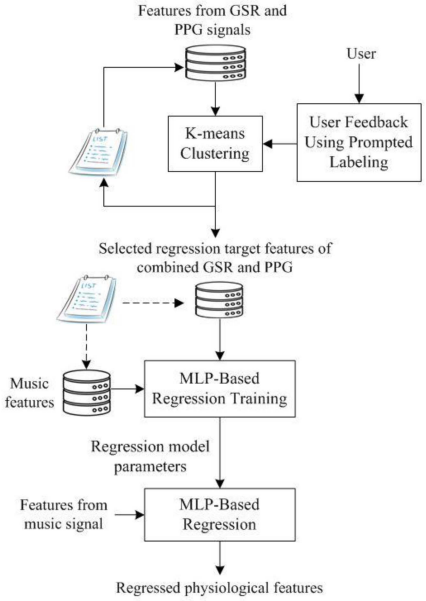

Fig. 4. Regression target feature selection and automatic generation of regression physiological features.

C. Automatic Generation of Regression Physiological Features

Human emotional changes caused by external stimuli generate physiological signals representing an internal state change within the body. Based on this phenomenon, features of the physiological signals corresponding to music can be automatically generated after the machine learns the features of physiological signals generated from humans when listening to music. To this end, we apply a regression method that learns the correlation between features extracted from physiological signals and musical features extracted from musical signals.

For regression training, we use a simple structured MLP with one hidden layer. In regression learning using the musical features as independent variables and the fused GSR and PPG features as dependent variables, the features extracted from GSR and PPG signals plays an important role as target values. To improve the classification accuracy, it is important to improve the training data into high-quality data. To this end, as shown in Fig. 4, we apply a prompted labeling technique to select the regression target features and apply a method of automatically generating physiological features through a regression model. The step-by-step process is as follows. (*Step* 1) By applying the features extracted from the training data composed of the GSR and PPG signals to the K-means clustering, we group them into n emotion clusters and form the models; (*Step* 2) The emotions are automatically classified by inputting the GSR and PPG signals (generated when listening to music) to the hierarchical inner attention-based deep neural network; (*Step* 3) The prompted labeling technique is applied to the classified emotion results. Through the GUI, a question about the match/disagreement of the perceived result to the user is transmitted through an alarm message, and immediate feedback is provided to the user. If the classified emotion result matches the user's feedback, the similarity between the

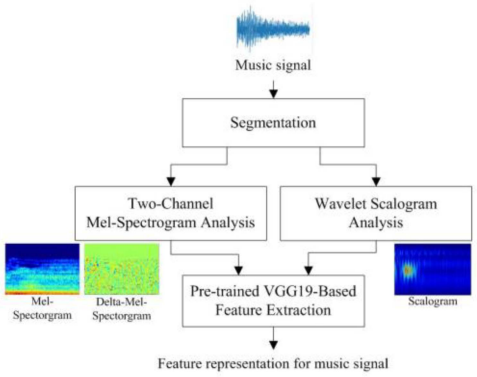

Fig. 5. Musical feature extraction from music signals.

features of the corresponding physiological signals and the cluster models created through the training data is compared by applying the *k*-means clustering; (*Step* 4) If the label fed back by the user and the label of the cluster model are similar, the music signals of the corresponding features are added to the training data, or deleted otherwise. Thus, high-quality features corresponding to *n* emotion clusters are stored in the new training database. (*Step* 5) the scalogram sample-level inner attention-mechanism-based CBGRU method is applied to the new training database to form optimal models suitable for each emotion. The features extracted through the retrained models are selected as target values for MLP-based regression learning, and consistently provide robust personalized physiological features according to user feedback. (*Step* 6) The weights trained through regression are applied to MLP-based regressors, and in the application phase, musical features extracted from music signals are inputted into the regressor. The fused GSR and PPG features mapped to musical features are automatically generated via the regressor without the need for GSR and PPG sensors. The generated features are applied to the BGRU classifier with segment-level inner attention mechanism.

D. Musical Feature Extraction From Music Signals

Fig. 5 illustrates the encoder for the musical feature extraction of the proposed method.

First, the input music signal is divided into segments of 0.5 s. When dividing the music, a 30% overlap between segments is used. Each music segment is converted and combined into a two-dimensional visual temporal-frequency representation, such as a Mel-spectrogram and a wavelet scalogram, before being input into a convolutional neural network that is pretrained through transfer learning. We use 64 Mel-filter banks for the spectrum of the music signal and add the energy values to extract the Mel-spectrum. The scalogram can enlarge part of the spectrogram, and the music signal is analyzed using 12 filter banks of equal width in the octave band.

To extract the low-dimensional deep spectral auditory (musical) features from the scalograms and Mel-spectrograms of the music signals, we use the VGG19 convolutional neural network (CNN) architecture [49] as a pretrained convolutional neural network. VGG19 consists of 19 layers: 16 convolutional layers, 3 fully connected layers, 5 maximum pooling layers, and 1 softmax layer. In this arrangement, we use a trained convolutional base with a densely connected classifier for normalization, and include a dropout layer.

As VGG19 requires raw image data as input, it extracts the static and first differential channels of the Mel-spectrogram segment and applies them as appropriate inputs to VGG19, along with the third channel representing the scalogram segment. These three channels can be regarded as an RGB image representation of audio data. The VGG19 provides 512 low-dimensional musical features.

III. Experiments and Results

A. Evaluation Data Sets and Experimental Methods

To evaluate the performance of the proposed method, we used two databases, including the Human Emotional Biosignal Database for Music (M-HEPS), and DEAP [50].

M-HEPS is a database that contains physiological data measured from GSR and PPG sensors of 18 participants who listened to music collected for the experiment. We selected 510 songs, 130 of which were classified as high arousal and high valence (HAHV), 123 as high arousal and low valence (HALV), 132 as low arousal and high valence (LAHV), and 125 as low arousal and low valence (LALV). Each song was less than 3 min long. All songs were encoded as standard MP3 audio files in stereo at 44.1 kHz and 128 kbps. For the measurement of GSR and PPG signals, music segments of about 40 s that clearly reflected four emotions were selected. Physiological data collection was carried out using PPG and GSR sensors with a sampling frequency of 128 Hz.

The DEAP dataset [50] consists of the PPG and GSR of 32 participants of various ages between 19 and 37 years old, as well as various physiological signals. Each participant watched the highlights of 40 pre-selected 1-minute music videos aimed to elicit emotions. All subjects then assigned each video a rating value from 1 to 9 in terms of arousal, valence, dominance, enjoyment, and familiarity. Arousal and valence levels ranged from 1–9, and hence arousal and valence were divided into two binary classes according to a threshold of 5 (high/low arousal and valence).

All the classification results in this study were calculated using a 10-fold cross-validation evaluation. Several experiments were conducted to measure the performance of the proposed method. The first aimed to classify emotions using only GSR and PPG signals that occur while listening to music. To this end, experiments were conducted using two sets of data. To compare the performance of the proposed and existing methods, we conducted cross-target emotion classification experiments at arousal and valence levels using an open DEAP data set. Furthermore, using the M-HEPS data set, the categorical recognition of emotions through GSR and PPG signals was evaluated on four emotion classes (HAHV, HALV, LAHV, and LALV). In the experiment, we tested several classification methods that are listed below.

1) RS-BF + RF [44]: This method extracts 22 features (BF) from raw signals (RS) measured from GSR and PPG sensors, which are then applied to the Random Forrest (RF)-based classification method.
2) PSD + SVM [51]: This method extracts power spectral density (PSD) from GSR and PPG signals, which are then applied to the support vector machine (SVM)-based classification method.
3) RS + HA-BGRU [52]: Raw sequences measured from the GSR and PPG sensors are combined with a traditional hierarchical attention-mechanism-based BGRU method (HA-BGRU). The features are extracted from the GSR and PPG sequences through the sample-level attention-mechanism-based BGRU encoder and applied to the segment-level attention- mechanism-based BGRU classifier.
4) RS + HA-CBGRU: Instead of HA-BGRU, we applied a classifier based on HA-CBGRU. In the HA-CBGRU method, features are extracted from GSR and PPG sequences through a sample-level attention- mechanism-based CBGRU encoder, and emotion classification is performed through a segment-level attention-mechanism -based BGRU classifier.
5) RS + IA-BLSTM [47]: Instead of HA-CBGRU, we applied a classifier based on bidirectional long short-term memory (BLSTM) model with inner attention mechanism. Once the segment vectors are output from BLSTM, a softmax layer is used over the output for classification.
6) SG + HA-CBGRU: This method extracts the scalogram from the raw signals measured by the GSR and PPG sensors, and the extracted scalogram is applied to an HA-CBGRU based classifier.
7) SG + HIA-CBGRU: Instead of HA-CBGRU, we applied the scalogram to a classifier based on HIA-CBGRU. In the HIA-CBGRU method, an inner attention mechanism is applied to the CBGRU encoder and BGRU classifier. This is our proposed method.

In the second experiment, to examine the relationship between segment length, which is divided from GSR and PPG signals, and the performance of emotion recognition, the performance was measured by dividing each GSR and PPG sequence into 0.5 s, 1 s, 2.5 s, and 5 s segments, respectively.

In the third experiment, the effect of the musical features extracted by the proposed method on emotion classification through regression training was compared with the following three baseline methods and two state-of-the-art methods.

1) Li *et al.* [53]: A deep convolutional neural network method on the music spectrograms that contains both the original time and frequency domain information is used.
2) Malik *et al.* [54]: A convolutional recurrent neural network method on long mel-band energy feature is applied to predict emotional response produced by music.
3) SP + MLP + SIA-BGRU: Musical features were extracted from the spectrum of music signals and MLP-based regression was used to learn the relationship between the fused features of the GSR and PPG signals and musical features extracted from the music signals. Based on the learned regression model, the fused features of the GSR and PPG signals for the input music were automatically generated and applied to the SIA-BGRU classifier.
4) MS + MLP + SIA-BGRU: Instead of SP, musical features were extracted by converting music signals to the Mel- spectrum (MS). The extracted features were applied to the MLP regressor and SIA-BGRU classifier.
5) SG + MLP + SIA-BGRU: Instead of MS, the musical features were extracted after converting the music signals to a scalogram (SG). The extracted features were applied to the MLP regressor and SIA-BGRU classifier.
6) MS-SG + SIA-BGRU: MS and SG were combined and applied to the MLP regressor and the SIA-BGRU classifier. This is our proposed method.

TABLE I
EMOTION CLASSIFICATION RESULTS BASED ON FUSION OF GSR AND PPG SIGNALS (%)

Features + Methods	Recognition Accuracy		
	DEAP		M-HEPS
	Valence	Arousal	
RS-BF + RF [44]	70.28	71.23	74.78
PSD + SVM [51]	71.34	72.15	75.46
RS + IA-BLSTM [47]	80.52	78.04	82.63
RS + HA-BGRU [52]	84.35	82.16	86.36
RS + HA-CBGRU	87.41	85.12	89.52
SG + HA-CBGRU	90.32	88.13	93.05
SG + HIA-CBGRU	92.41	90.27	95.32

B. Experimental Results

Table I presents the experimental results of GSR and PPG signal-based emotion classification using the proposed method, as well as comparisons with several methods that include other types of neural network architectures. The obtained values are the total averages of the emotion recognition results evaluated for each individual. In the case of the emotion classification experiment that used only GSR and PPG signals, a total of 450 GSR and PPG clips that clearly show the emotional state were used, and each clip lasted for about 40 s. Each 40 s long GSR and PPG sequence was divided into 0.5 s segments to be applied to the hierarchical inner attention mechanism.

As shown in Table I, the best results for both databases were achieved using the proposed method, SG + HIA-CBGRU, and the classification accuracy of SG + HA-CBGRU was slightly lower than that of the proposed method. From this result, it may be observed that the inner attention mechanism performed better than the attention mechanism. Further, the results of these two methods were superior to those obtained with the other five methods. Therefore, it may be observed that using the characteristics obtained through scalogram analysis rather than the raw data produced improved results. In addition, as a method of extracting features from GSR and PPG sequences, the use of the CBGRU encoder was more effective than the

TABLE II
EMOTIONAL CLASSIFICATION ACCURACY ON DIFFERENT LENGTH SEGMENTS FOR DEAP DATA SET (%)

Features + Methods	0.5s	1s	2.5s	5s
PSD + SVM [51]	71.34	69.98	68.23	65.87
RS + HA-BGRU [52]	84.35	83.04	81.29	79.08
RS + HA-CBGRU	87.41	86.12	84.32	82.35
SG + HA-CBGRU	90.32	88.96	86.98	84.35
SG + HIA-CBGRU	92.41	91.06	89.24	86.93

TABLE III
MUSIC CLASSIFICATION RESULTS USING REGRESSION MODEL (%)

Features + Methods	Recognition Accuracy	
	Individual	Common
Li *et al.* [53]	83.23	74.85
Du *et al.* [54]	87.43	78.92
SP + MLP + SIA-BGRU	91.51	83.18
MS + MLP + SIA-BGRU	93.18	85.33
SG + MLP + SIA-BGRU	94.23	86.64
MS-SG + MLP + SIA-BGRU	96.26	88.75

use of the BGRU encoder. Additionally, emotion classification results for M-HEPS were higher than those for DEAP. This is attributed to M-HEPS containing more emotionally distinct music than DEAP. Compared to the other five methods, the method using RS-BF + RF attained the worst results.

To further explore the relationship between the performance of the proposed model and the segment length of the GSR and PPG sequence, a valence classification experiment using five models was performed. The experimental results from the DEAP data set are shown in Table II. The classification accuracy of the five models is continuously improved by reducing the length of the segment. Therefore, the SG + HIA-CBGRU model achieved the best cross-validation classification accuracy of 92.41% with 0.5 s segmented GSR and PPG sequences, which was 2.09%, 5.00%, 8.06%, and 21.07% higher than that of SG + HA-CBGRU, RS + HA-CBGRU, RS + HA-BGRU, and PSD + SVM, respectively.

When the GSR and PPG sequences were split into 0.5 s segments, one sequence is divided into 120 segments, and each segment contained 64 samples. Non-stationarity was not evident because the segment had a short range. Using this approach, all four models provided their optimal performance with 0.5 s segmented GSR and PPG sequences. In particular, the hierarchical structure and inner attention mechanism play a significant role in the proposed method in improving its performance. From the fourth rows in Table II, we can observe that the accuracy of SG + HIA-CBGRU on 0.5 s segmented sequences is 1.35%, 3.17%, and 5.48% higher than that on 1 s, 2.5 s, and 5 s segmented sequences, respectively.

Table III summarizes the results of applying various musical feature extraction methods to the same MLP regression and SIA-BGRU classifier for music classification. We tested two different approaches for selecting the fused feature vectors from GSR and PPG signals related to each emotion. The first was the individual approach, which individually applied the fusion feature vectors extracted from each individual's GSR and PPG signals for regression learning. The second was the common, subject-independent approach that applied fusion feature vectors obtained by learning the GSR and PPG signals from all the participants in each class of emotion. Music classification performance was evaluated for testing the partitioned music of data sets that were never used in encoders for GSR and PPG or MLP-based regression. Our method MS-SG + MLP + SIA-BGRU achieved the best cross-subject classification accuracy of 96.26%, which was 2.03%, 3.08%, and 4.75% higher than SG + MLP + SIA-BGRU, MS + MLP + SIA-BGRU, and SP + MLP + SIA-BGRU respectively in the case of "individuals." In addition, it was confirmed that the application of prompted labeling and detailed musical feature extraction improved the classification accuracy of the method MS-SG + MLP + SIA-BGRU more than that of the method SG + MLP + SIA-BGRU. As expected, our method outperformed two state-of-the-art methods [53], [54] which only use content-based deep learning representations.

We installed an emotion-based music classification system on a smartphone (including an Octa-Core 2.7GHz CPU and 12 GB RAM). Owing to the limited resources of the smartphone, we applied 8-bit quantization to reduce the size of the regression and classification model trained on the server before applying it to mobile devices. The implemented system divided the received music into 0.5 s segments and processes them in parallel. The system's CPU usage was 29.31%, and the memory size was 35.04 MB. In addition, when the system was maintained for 30 min, a power consumption of 2.1573 mAh/m was confirmed. End-to-end prefetching is applied to reduce the latency of the algorithm to 1240 ms for an input music file with a length of 40 s. The prediction accuracy of music classification increased because it used a multi-layer architecture that plays an important role for each layer. However, a method that can reduce the amount of computation is still required because more time is needed for learning and inference.

IV. CONCLUSION

In this study, we proposed a system for classifying music through categorical recognition of emotions using GSR and PPG signals induced by music. The system includes two methods. The first method predicted the user's emotions with high accuracy by applying the data acquired through wearable GSR and PPG sensors to the CBGRU method with a hierarchical inner attention mechanism. Additionally, the second method learned the relationship between emotion-specific features extracted from previously generated GSR and PPG signals and musical features extracted from music. Then, fusion features of GSR and PPG signals suitable for music were automatically generated based on their regression learning capabilities and applied to SIA-BGRU classifiers. We believe the results obtained in this study help to effectively interpret the response of human physiological signals to musical stimuli and relate them to machine learning. In future research, we will conduct a more detailed emotion classification experiment using the DEAP dataset with values from 1 to 9, and develop a

deep learning approach that classifies more emotion classes using various physiological signals such as GSR, PPG, EEG, and ECG.

REFERENCES

[1] M. Schedl, H. Zamani, C.-W. Chen, Y. Deldjoo, and M. Elahi, "Current challenges and visions in music recommender systems research," *Int. J. Multimedia Inf. Retrieval*, vol. 7, no. 2, pp. 95–116, 2018, doi: 10.1007/s13735-018-0154-2.

[2] S.-C. Lim, J.-S. Lee, S.-J. Jang, S.-P. Lee, and M. Y. Kim, "Music-genre classification system based on spectro-temporal features and feature selection," *IEEE Trans. Consum. Electron.*, vol. 58, no. 4, pp. 1262–1268, Nov. 2012, doi: 10.1109/TCE.2012.6414994.

[3] C.-H. Lee, J.-L. Shih, K.-M. Yu, and H.-S. Lin, "Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features," *IEEE Trans. Multimedia*, vol. 11, no. 4, pp. 670–682, Jun. 2009, doi: 10.1109/TMM.2009.2017635.

[4] S. Shin, D. Jang, J. J. Lee, S.-J. Jang, and J.-H. Kim, "MyMusicShuffler: Mood-based music recommendation with the practical usage of brain-wave signals," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2014, pp. 355–356, doi: 10.1109/ICCE.2014.6776039.

[5] S.-H. Lee, T.-Y. Chen, Y.-T. Hsien, and L.-R. Cao, "A music recommendation system for depression therapy based on EEG," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2020, pp. 1–2, doi: 10.1109/ICCE-Taiwan49838.2020.9258021.

[6] S. Shirali-Shahreza, H. Abolhassani, and M. H. Shirali-Shahreza, "Fast and scalable system for automatic artist identification," *IEEE Trans. Consum. Electron.*, vol. 55, no. 3, pp. 1731–1737, Aug. 2009, doi: 10.1109/TCE.2009.5278049.

[7] K. Nakamura, T. Fujisawa, and T. Kyoudou, "Music recommendation system using lyric network," in *Proc. IEEE Global Conf. Consum. Electron.*, 2017, pp. 1–2, doi: 10.1109/GCCE.2017.8229316.

[8] S. Hong, K. Y. Lee, and K. Y. Lee, "Fast and adaptive browsing state recovery for multimedia consumer electronics devices," *IEEE Trans. Consum. Electron.*, vol. 57, no. 1, pp. 164–172, Feb. 2011, doi: 10.1109/TCE.2011.5735498.

[9] A. Baijal, V. Agarwal, and D. Hyun, "Analyzing images for music recommendation," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2021, pp. 1–6, doi: 10.1109/ICCE50685.2021.9427619.

[10] J. Lee, S. Shin, D. Jang, S.-J. Jang, and K. Yoon, "Music recommendation system based on usage history and automatic genre classification," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2015, pp. 134–135, doi: 10.1109/ICCE.2015.7066352.

[11] W.-I. Park, S. Kang, M. Choi, and Y.-K. Kim, "A music recommendation system in mobile environment," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2009, pp. 1–2, doi: 10.1109/ICCE.2009.5012299.

[12] J. Lee and C. S. Hong, "Personal multimedia recommendation system on smartphone platform," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2012, pp. 586–587, doi: 10.1109/ICCE.2012.6161984.

[13] K. Yoon, J. Lee, and M.-U. Kim, "Music recommendation system using emotion triggering low-level features," *IEEE Trans. Consum. Electron.*, vol. 58, no. 2, pp. 612–618, May 2012, doi: 10.1109/TCE.2012.6227467.

[14] T. Maekaku and H. Kasai, "Music relationship visualization based on melody piece transition using conditional divergence," *IEEE Trans. Consum. Electron.*, vol. 58, no. 3, pp. 1006–1012, Aug. 2012, doi: 10.1109/TCE.2012.6311349.

[15] X . Zhu, Y.-Y. Shi, H.-G. Kim, and K.-W. Eom, "An integrated music recommendation system," *IEEE Trans. Consum. Electron.*, vol. 52, no. 3, pp. 917–915, Aug. 2006, doi: 10.1109/TCE.2006.1706489.

[16] Y. Kodama *et al.*, "A music recommendation system," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2005, pp. 219–220, doi: 10.1109/ICCE.2005.1429796.

[17] T. Maekaku and H. Kasai, "Music classification applying prime form and interval-class vector," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2013, pp. 312–313, doi: 10.1109/ICCE.2013.6486906.

[18] R. L. Rosa, D. Z. Rodriguez, and G. Bressan, "Music recommendation system based on user's sentiments extracted from social networks," *IEEE Trans. Consum. Electron.*, vol. 61, no. 3, pp. 359–367, Aug. 2015, doi: 10.1109/TCE.2015.7298296.

[19] H.-J. Won, W.-J. Yoon, and K.-S. Park, "P2P music recommender system based on a single-scaled hybrid filtering," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2011, pp. 817–818, doi: 10.1109/ICCE.2011.5722881.

[20] F. Fessahaye *et al.*, "T-RECSYS: A novel music recommendation system using deep learning," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2019, pp. 1–6, doi: 10.1109/ICCE.2019.8662028.

[21] D. Sánchez-Moreno, A. B. G. González, M. D. M. Vicente, V. F. L. Batista, and M. N. M. García, "A collaborative filtering method for music recommendation using playing coefficients for artists and users," *Expert Syst. Appl.*, vol. 66, pp. 234–244, Dec. 2016, doi: 10.1016/j.eswa.2016.09.019.

[22] A. Van den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2643–2651.

[23] X. Wang, D. Rosenblum, and Y. Wang, "Context-aware mobile music recommendation for daily activities," in *Proc. ACM Int. Conf. Multimedia*, 2012, pp. 99–108, doi: 10.1145/2393347.2393368.

[24] H.-G. Kim, G. Y. Kim, and J. Y. Kim, "Music recommendation system using human activity recognition from accelerometer data," *IEEE Trans. Consum. Electron.*, vol. 65, no. 3, pp. 349–358, Aug. 2019, doi: 10.1109/TCE.2019.2924177.

[25] H. Liu, J. Hu, and M. Rauterberg, "LsM: A new location and emotion aware web-based interactive music system," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2010, pp. 253–254, doi: 10.1109/ICCE.2010.5418750.

[26] N.-H. Liu and H.-Y. Kung, "JoMP: A mobile music player agent for joggers based on user interest and pace," *IEEE Trans. Consum. Electron.*, vol. 55, no. 4, pp. 2225–2233, Nov. 2009, doi: 10.1109/TCE.2009.5373792.

[27] P. Lisena, R. Troncy, K, Todorov, and M. Achichi, "Modeling the complexity of music metadata in semantic graphs for exploration and discovery," in *Proc. 4th Int. Digit. Libraries Musicol. Workshop*, 2017, pp. 17–24, doi: 10.1145/3144749.3144754.

[28] Y.-C. Yu, S. D. You, and D.-R. Tsai, "Magic mirror table for social-emotion alleviation in the smart home," *IEEE Trans. Consum. Electron.*, vol. 58, no. 1, pp. 126–131, Feb. 2012, doi: 10.1109/TCE.2012.6170064.

[29] S. Bianco, L. Celona, and P. Napoletano, "Visual-based sentiment logging in magic smart mirrors," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2018, pp. 1–4, doi: 10.1109/ICCE-Berlin.2018.8576217.

[30] S. Bazrafkan, T. Nedelcu, P. Filipczuk, and P. Corcoran, "Deep learning for facial expression recognition: A step closer to a smartphone that knows your moods," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2017, pp. 217–220, doi: 10.1109/ICCE.2017.7889290.

[31] K. Sakuurai, R. Togo, T. Ogawa, and M. Haseyama, "Music playlist generation based on reinforcement learning using acoustic feature map," in *Proc. IEEE Global Conf. Consum. Electron.*, 2020, pp. 942–943, doi: 10.1109/GCCE50665.2020.9291748.

[32] J. C. Yu, I.-C. Chang, and Y.-J. Lin, "A dynamic music adding system based on cloud computing," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2015, pp. 262–263, doi: 10.1109/ICCE-TW.2015.7216888.

[33] Y.-T. Wan, C.-C. Chiu, K.-W. Liang, and P.-C. Chang, "Midoriko chatbot: LSTM-based emotional 3D avatar," in *Proc. IEEE Global Conf. Consum. Electron.*, 2019, pp. 937–940, doi: 10.1109/GCCE46687.2019.9015303.

[34] R. L. Rosa, D. Z. Rodrigues, G. M. Schwartz, I. C. Ribeiro, and G. Bressan, "Monitoring system for potential users with depression using sentiment analysis," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2016, pp. 381–382, doi: 10.1109/ICCE.2016.7430656.

[35] L. Papa, A. Sabatelli, L. Ciabattoni, A. Monteriu, F, Lamberti, and L. Morra, "Stress detection in computer users from keyboard and mouse dynamics," *IEEE Trans. Consum. Electron.*, vol. 67, no. 1, pp. 12–19, Feb. 2021, doi: 10.1109/TCE.2020.3045228.

[36] A. C. Generosi, S. Ceccacci, and M. Mengoni, "A deep learning-based system to track and analyze customer behavior in retail store," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2018, pp. 1–6, doi: 10.1109/ICCE-Berlin.2018.8576169.

[37] B. Fong and J. Westerink, "Affective computing in consumer electronics," *IEEE Trans. Affective Comput.*, vol. 3, no. 2, pp. 129–131, Apr.–Jun. 2012, doi: 10.1109/T-AFFC.2012.20.

[38] D. K. Kim, S. Ahn, S. Park, and M. Whang, "Interactive emotional lighting system using physiological signals," *IEEE Trans. Consum. Electron.*, vol. 59, no. 4, pp. 765–771, Nov. 2013, doi: 10.1109/TCE.2013.6689687.

[39] C.-L Lin, P.-S. Gau, K.-J. Lai, Y.-K. Chu, and C.-H. Chen, "Emotion Caster: Tangible emotion sharing device and multimedia display platform for intuitive interactions," in *Proc. IEEE Int. Symp. Consum. Electron.*, 2009, pp. 988–989, doi: 10.1109/ISCE.2009.5156954.

[40] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, Sep. 2017, doi: 10.1016/j.inffus.2017.02.003.

[41] R. Chatterjee, S. Mazumdar, R. S. Sherratt, R. Halder, T. Maitra, and D. Giri, "Real-time speech emotion analysis for smart home assistants," *IEEE Trans. Consum. Electron.*, vol. 67, no. 1, pp. 68–76, Feb. 2021, doi: 10.1109/TCE.2021.3056421.
[42] I. Lee, H. Jung, C. Ahn, J. Seo, J. Kim, and O. Kwon, "Real-time personalized facial expression recognition system based on deep learning," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2016, pp. 267–268, doi: 10.1109/ICCE.2016.7430609.
[43] J. Shin, J. Maeng, and D.-H. Kim, "Inner emotion recognition using multi bio-signals," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2018, pp. 206–212, doi: 10.1109/ICCE-ASIA.2018.8552152.
[44] D. Ayata, Y. Yaslan, and M. E. Kamasak, "Emotion based music recommendation system using wearable physiological sensors," *IEEE Trans. Consum. Electron.*, vol. 64, no. 2, pp. 196–203, May 2018, doi: 10.1109/TCE.2018.2844736.
[45] Y.-K. Lee, O.-W. Kwon, H. S. Shin, J. Jo, and Y. Lee, "Noise reduction of PPG signals using a particle filter for robust emotion recognition," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2011, pp. 202–205, doi: 10.1109/ICCE-Berlin.2011.6031807.
[46] A. Abdul, J. Chen, H.-Y. Liao, and S.-H. Chang, "An emotion-aware personalized music recommendation system using a convolutional neural networks approach," *Appl. Sci.*, vol. 8, no. 7, p. 1103, 2018, doi: 10.3390/app8071103.
[47] Y. Liu, C. Sun, L. Lin, and X. Wang, "Learning natural language inference using bidirectional LSTM model and inner-attention," 2016. [Online]. Available: arXiv:1605.09090.
[48] Z. Ren, K. Qian, Z. Zhang, V. Pandit, A. Baird, and B. Schuller, "Deep scalogram representations for acoustic scene classification," *IEEE/CAA J. Automatica Sinica*, vol. 5, no. 3, pp. 662–669, 2018, doi: 10.1109/JAS.2018.7511066.
[49] I.-J. Ding and N.-W. Zheng, "Classification of restlessness level by deep learning of visual geometry group convolution neural network with acoustic speech and visual face sensor data for smart care applications," *Sens. Mater.*, vol. 32, no. 7, pp. 2329–2341, 2020, doi: 10.18494/SAM.2020.2881.
[50] S. Koelstra *et al.*, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 18–31, Jan.–Mar. 2012, doi: 10.1109/T-AFFC.2011.15.
[51] G. Udovičić, J. Derek, M. Russo, and M. Sikora, "Wearable emotion recognition system based on GSR and PPG signals," in *Proc. 2nd Int. Workshop Multimedia Pers. Health Health Care*, 2017, pp. 53–59.
[52] J. X. Chen, D. M. Jiang, and Y. N. Zhang, "A hierarchical bidirectional GRU model with attention for EEG-based emotion classification," *IEEE Access*, vol. 7, pp. 118530–118540, 2019, doi: 10.1109/ACCESS.2019.2936817.
[53] X. Li, J. Tian, M. Xu, Y. Ning, and L. Cai, "DBLSTM based multi-scale fusion for dynamic emotion prediction in music," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2016, pp. 1–6, doi: 10.1109/ICME.2016.7552956.
[54] M. Malik, S. Adavanne, K. Drossos, T. Virtanen, D. Ticha, and R. Jarina, "Stacked convolutional and recurrent neural networks for music emotion recognition," in *Proc. 14th Sound Music Comput. Conf.*, 2017, pp. 208–213.

Hyoung-Gook Kim received the Dr.-Ing. degree in electrical engineering and computer science from the Technical University of Berlin, Germany. Since 2007, he has been a Professor with the Department of Electronic Convergence Engineering, Kwangwoon University, Seoul, Republic of Korea. His research interests include audiovisual content indexing and retrieval, biomedical signal processing, and deep learning.

Gi Yong Lee received the B.S. degree in electronic convergence engineering from Kwangwoon University, Seoul, Republic of Korea, in 2020, where he is currently pursuing the master's degree. His research interest includes machine learning.

Min-Soo Kim received the B.S. degree in electronic convergence engineering from Kwangwoon University, Seoul, Republic of Korea, in 2021, where he is currently pursuing the master's degree. His research interest includes audio signal processing.