IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 July 2023, accepted 24 July 2023, date of publication 31 July 2023, date of current version 3 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3300042

RESEARCH ARTICLE

Musi-ABC for Predicting Musical Emotions

JING YANG
Conservatory of Music, Qilu Normal University, Jinan 250200, China
e-mail: ttv8r3@sina.com

ABSTRACT To address the issues of insufficient accuracy and low training efficiency in general musical emotion prediction models, we propose the muSi-ABC architecture for predicting music emotions. Specifically, in the feature extraction stage of music emotions, we use a benchmark feature set to ensure that the extracted music emotion features adhere to standardization. In the prediction stage, we introduce the muSi-ABC architecture which first utilizes a 2D-ConvNet (two dimensional-Convolutional Neural network) to extract partial critical features in music emotions. Then, the BiLSTM (Bi-directional Long Short Term Memory) neural network is employed to learn contextual sequential information of past and future music emotions from the obtained partial critical features. Furthermore, the SA (Self-Attention) module is applied to obtain the complete critical features highly relevant to music emotions, thereby improving prediction accuracy and training efficiency. Through ablation experiments conducted at different time term lengths, the roles of ConvNet model and SA module, as well as the advantages of the proposed muSi-ABC architecture over other ablated models in terms of training efficiency and prediction accuracy, are verified. Additionally, it is observed that representing music emotions using long term feature information for the same song can enhance prediction accuracy. Finally, contrast experimental results demonstrate that the proposed architecture outperforms other benchmark methods in terms of prediction accuracy. Moreover, it is validated that the outlier points contained in the music emotions features extracted based on the benchmark feature set help discover the variations trends of music emotions.

INDEX TERMS Predicting musical emotions, long term dependency, partial critical features, complete critical points.

I. INTRODUCTION

Music is a language that can express emotions, specifically, the composers and performers express their inner emotions through music, and the listeners resonate with the emotions expressed in the music, leading to an understanding of the emotional essence of the music. Musical emotions are the subjective description of one's inner psychological state while listening to music, which is influenced by a combination of internal subjective factors and external objective factors [1], [31]. Musical emotions evolve over time as the melody, harmony, and rhythm of the music change, and they encompass subjectivity and complexity, as well as the temporal and continuous nature of music. The emotional features of music are complex and diverse, providing listeners with rich emotions. While humans have the ability to perceive the rich emotions in music, computers are still unable to do so. Therefore, predicting the emotions expressed in music poses a great challenge for computers.

Computers attempt to develop the ability to predict musical emotions like humans by intelligent computation [32]. Specifically, by using neural networks, the computer can analyze the features in music emotions that are input into the model, thus predicting the music emotions. At present, the prediction network mainly analyzes the input musical emotion features through the recurrent neural network (RNN) and identifies the music emotions. In addition, different time slices of a song represent different emotional forms, and in order to find the critical information representing musical emotions in a slice, the convolutional neural network (ConvNet) is introduced into the RNN to effectively capture critical musical emotion information within partial time slices [2]. Furthermore, there is a different correlation between the musical emotion feature information contained in different

The associate editor coordinating the review of this manuscript and approving it for publication was Luca Turchet.

time slices of the song and the musical emotion. In order to capture the most relevant feature information related to musical emotions, the attention module [3] is introduced into the neural network, which can effectively capture the feature information most relevant to the musical emotion in the complete data and thereby improve the accuracy of emotions predicting. Recent related studies focus on the design of network models based on RNN, with emphasis on the impact of partial and complete critical information on musical emotions. Additionally, most studies verify the performance of the model by predicting the musical emotions of labeled songs [33], [34].

In practical applications, people predict the main melody of a song through the auditory system. By considering the relevance of the context and combining the emotional information obtained with the stored musical emotion memory in the brain, they analyze the complete critical musical emotion information. This process allows humans to predict the emotions expressed in music. Taking inspiration from this, we propose the muSi-ABC architecture, which combines ConvNet, BiLSTM (Bidirectional Long Short-Term Memory), and SA (Self-Attention) models, to simulate the process of predicting musical emotions similar to humans. Specifically, based on ConvNet, the proposed method extracts partial critical features of musical emotion, uses BiLSTM neural network to learn the context sequences of musical emotions past and future from extracted partial critical features, and introduces SA mechanism to obtain complete critical features information highly relevant to musical emotions. Finally, The contrast and ablation experimental results validate the effectiveness of the proposed method.

II. RELATED WORK

In studying tasks about musical emotions prediction, the existent models can be divided into two categories: traditional machine learning methods and deep learning methods [35].

Most traditional machine learning methods for predicting musical emotions are statistical probability models. The selection and combination of handcrafted features have a significant impact on the learning effectiveness of the model, making them suitable for handling the limited-sample problems. Initially, researchers often used Support Vector Machines (SVM) or combined that with other statistical probability models to classify musical emotions. Although they achieved good prediction results, there is uncertainty in the emotional classification criteria. To address this issue, Cai et al. [4] first introduced using regression training to solve the music emotions prediction problem. They concatenated the features extracted from different feature tools into 114-dimensional musical features and used Support Vector Regression (SVR) models to identify the Valence and the Arousal of each music sample. Xiang et al [5] used seven different music features to identify continuous dimensional emotional values based on the SVR model and compared it with the SVM model. The experimental results showed that SVR performed better than SVM in predicting dimensional emotions.

In recent years, with the development of deep learning, the accuracy of using deep learning methods to predicting musical emotions has greatly improved [6]. Most deep learning music emotion prediction methods are based on neural network models. The design of the network model affects the prediction accuracy, making it suitable for handling large-sample data problems. The most commonly used neural network models can be divided into three categories: A. RNNs, B. a combination of ConvNets and RNNs, and C. Neural networks with fused attention models.

A. RNNS

Huang et al. [7] incorporated psychoacoustic features into the ComPareE feature set and used LSTM-RNN to model longer term contextual information, capture the temporal emotion features, and predicting musical emotions. Dutta et al. [8] proposed a Deep Bidirectional Long Short-Term Memory Extreme Learning Machine (DBLSTM-ELM) model that combines extreme learning machine to fuse the prediction results of DBLSTM of music emotions with different time intervals, and obtain the final decision. RNNs have performed well in solving temporal problems, but they do not consider the influence of partial critical information on musical emotions. Meanwhile, LSTM is at risk of overfitting during the training phase, and there are issues with low training efficiency and long term dependence.

B. A COMBINATION OF CONVNETS AND RNNS

Naser and Saha [9] used two ConvNet-based L3-Net and VGGish models with the deep audio embedding method to aggregate high-dimensional spectrogram features for predicting musical emotions, considering the influence of partial critical information. However, ConvNets did not consider the temporality of musical emotions, so the use of the single ConvNet or RNN cannot solve the musical prediction problem well. Dang et al. [10] introduced a deep learning model that combines 2D-ConvNet and RNN to analyze spectrogram features for predicting musical emotions. Satayarak et al. [11] proposed a method that combines transfer learning and CRNN (Convolutional Recurrent Neural Network) to extract emotional features in both the time-frequency domains of spectrograms for speech emotion prediction. Liu et al. [12] introduced a Convolutional Long Short-Term Memory Deep Neural Network (CLDNN) that combines Mel-Frequency Cepstral Coefficients (MFCC) spectrograms and Mel filterbank energy spectrogram features on base of standard acoustic statistics for predicting musical emotions. To address the low training efficiency problem of LSTM, Hasanzadeh et al. [13] found that ConvNet can learn directly from input data in image recognition tasks, thereby reducing the parameter size of spatial structure information and improving training efficiency.

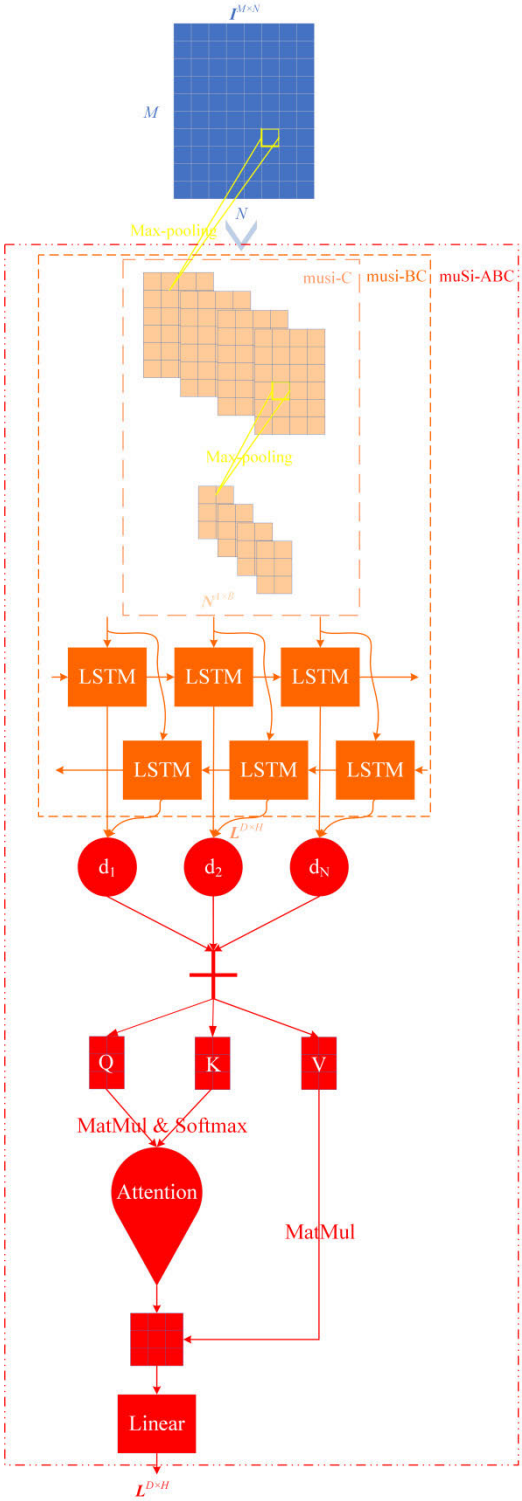

FIGURE 1. The proposed architecture. In the figure, musi-C is the musical emotion prediction model with ConvNet only, musi-BC is the musical emotion prediction model with ConvNet and BiLSTM, and muSi-ABC is the comprehensive model.

C. NEURAL NETWORKS WITH FUSED ATTENTION MODELS

To address the problem of long term dependence in LSTM, Huang et al. [14] proposed a hybrid LSTM model with attention mechanism to alleviate the reduction of learning contextual information with increasing input term over time in music. Traditional attention-based models rely heavily on external information, but the complex and diverse nature of musical emotions means that the overall emotional expression is not simply a simple summation of time and emotional features, but largely depends on the correlation with musical emotional features. To tackle this problem, Jiang et al. [3] introduced a Bidirectional Gate Recurrent Unit (BiGRU) network model with self-attention mechanism, for predicting musical emotions and themes. Compared with the hybrid LSTM model that integrates the traditional attention mechanism, the experimental results showed that the self-attention module exhibits stronger fitting ability and higher training efficiency than the traditional attention model.

In summary, considering the temporal and continuous nature of music emotions, BiLSTM is chosen as the basic model (referred to as musi-B) in this paper. To address the problem that LSTM do not consider the influence of partial critical information on musical emotions and have low training efficiency, a ConvNet-BiLSTM (in other words, musi-BC) model is constructed by integrating 2D-ConvNet. For the long term dependency problem of LSTM, a self-attention module is further integrated into the musi-BC model, forming the overall muSi-ABC architecture. By capturing partial critical information, sequential information, and complete critical information of musical emotions, the proposed considerate architecture addresses the limitations of LSTM in predicting long term musical emotions and improves training efficiency. Thus, it provides an effective method for enhancing the accuracy and efficiency of long term musical emotion prediction.

III. MATERIALS AND METHODS

A. OVERALL ARCHITECTURE AND FORMAL DEFINITION

The proposed muSi-ABC architecture comprises the two-dimensional convolutional layer, the bidirectional long short-term memory layer, and the self-attention layer (see Figure 1 for the overall model structure).

Firstly, each input song is represented as an $\boldsymbol{I}^{M\times N} = \{\boldsymbol{i}_1, i_{2,\ldots,} i_M\}$ of music emotional features, where M represents the time dimension and N represents the dimension of music emotional features. Furthermore, the output of musi-C (i.e., ConvNet) is denoted as $\boldsymbol{N}^{A\times B}$. Subsequently, the output of musi-BC is represented as $\boldsymbol{L}^{D\times H}$. Lastly, the holistic output of muSi-ABC is denoted by $\boldsymbol{A}^{V\times H}$.

B. BACKBONE MODEL

The proposed muSi-ABC architecture simulates the process of human music prediction and emotional expression. It utilizes the two-dimensional ConvNet model to extract melody slices, the BiLSTM network to obtain emotional context information, and the SA module to combine obtained emotional information with stored emotional memory, resulting in complete critical music emotion information.

1) 2D-ConvNet

To obtain partial critical features of musical emotions from the two dimensions of time and music emotion features in the feature matrix, a two-dimensional CNN is used for processing, as shown in Figure 2. Taking the prediction of the continuous emotional values of a song as an example, the musical emotion feature matrix, i.e., $I^{M\times N}$,is first input into the two-dimensional convolutional layer, which extracts music emotion features with a $K(3 \times 3)$ filter while preserving edge information. Then, BatchNorm2d is used for data normalization processing to ensure consistent distribution of the output data after convolution. Next, the ReLU activation function is used to add non-linear factors and enhance the ability of the two-dimensional convolutional layer to express music emotions. Finally, the maximum pooling (MaxPooling) method is selected to reduce the matrix dimension and preserve some critical information in the music emotion features, thus obtaining the feature matrix $N^{A\times B}$ about partial critical music emotions.

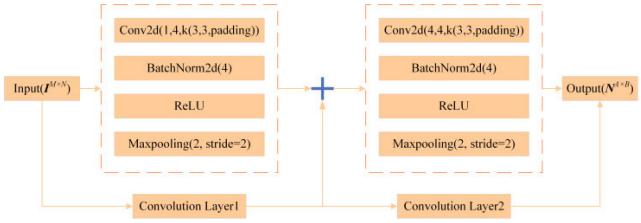

FIGURE 2. 2D-ConvNet.

2) BiLSTM

LSTM has a unidirectional transmission direction, from the previous time step to the next time step. However, music emotions have strong internal correlations, and the current state is not only related to the previous state but also to the next state. Therefore, the bidirectional LSTM network is constructed using two LSTM layers [15] to predict past and future emotional information in music and model the contextual information of music emotions.

The recurrent unit structure of LSTM includes three gates and two states, i.e., the input gate i_t, the forget gate f_t, the output gate o_t, the internal state c_t, and the candidate state c'_t, as shown in Figure 3. Assuming that the external state at time t is h_t and the external state at the previous time step is h_{t-1}. LSTM combines the previous external state h_{t-1} with the current input music emotion feature vector n_t. The three gate values and the candidate state value of the LSTM recurrent unit are calculated using (1)-(4). The memory unit c_t is updated using the forget gate f_t and the input gate i_t through (5), and the output gate o_t transfers the emotional information of the internal state to the external state h_t through (6).

$$i_t = \sigma(W_i n_t + U_i h_{t-1} + b_i) \quad (1)$$
$$f_t = \sigma(W_f n_t + U_f h_{t-1} + b_f) \quad (2)$$
$$o_t = \sigma(W_o n_t + U_o h_{t-1} + b_o) \quad (3)$$
$$c'_t = tanh(W_c n_t + U_c h_{t-1} + b_c) \quad (4)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot c'_t \quad (5)$$
$$h_t = o_t \odot tanh(c_t) \quad (6)$$

Here, $x \in \{i, f, o, c\}$ represents the components of W_x, U_x, and b_x, W_x is the weight matrix at the current time step, U_x is the weight matrix at the previous time step, and b_x is the bias vector, σ represents the sigmoid function, while *tanh* represents the hyperbolic tangent function.

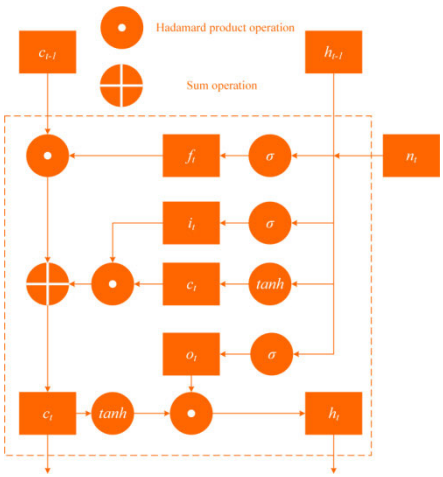

FIGURE 3. The structure of an LSTM recurrent unit.

The BiLSTM model consists of a forward layer and a backward layer of LSTM. (7) and (8) are utilized to extract and retain emotional information from both past and future music. Figure 4 illustrates the structure of the single-layer BiLSTM network. Assuming that the forward layer follows the time order while the backward layer follows the reverse time order, the hidden layer states at time t are defined as h_t^1 and h_t^2. The output vector $l_t$of the bidirectional Long Short-Term Memory layer at time t is computed based on the hidden layer states in both directions, as depicted in (9).

$$h_t^1 = f(U^1 h_{t-1}^1 + W^1 n_t + b^1) \quad (7)$$
$$h_t^2 = f(U^2 h_{t-1}^2 + W^2 n_t + b^2) \quad (8)$$
$$l_t = W^{t1} h_t^1 + W^{t2} h_t^2 + b^0) \quad (9)$$

In which, W_x $(x \in \{1, 2\})$ represents the weight matrix at the current time step, U^1 and U^2 represent the weight matrices at the previous and next time steps, f represents the activation function of the hidden layer, W^{tx} $(x \in \{1, 2\})$ represents the weight matrix of the hidden layer state at the current time step, and b^x $(x \in \{0, 1, 2\})$ represents the bias vector. After two layers of BiLSTM, a serialized music emotion feature matrix $L^{D\times H}$ is obtained.

3) SELF-ATTENTION

The musical emotion feature matrix $L^{D\times H}$, which represents the output of the bidirectional LSTM layer, is inputted to

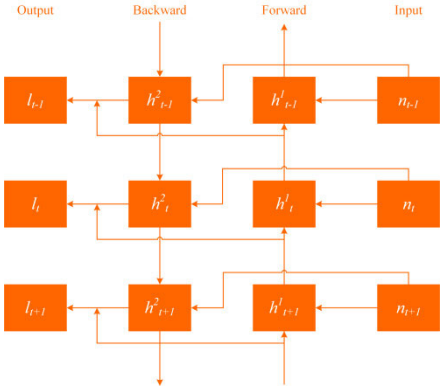

FIGURE 4. The structure of BiLSTM.

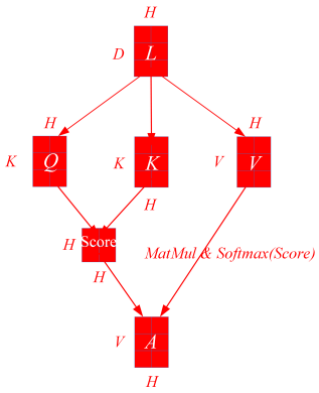

FIGURE 5. Self-attention module.

the self-attention layer. Each music emotion feature vector in matrix L at every time step is treated as a query vector and compared with the music emotion feature vectors at different time steps in the song to calculate similarity scores. After performing weighted averaging, the complete critical feature information about music emotions is obtained. The self-attention module structure, with the number of rows and columns labeled outside each box, is illustrated in Figure 5. The computation process is as follows

(1) For the input matrix L, the linear mapping is performed to obtain the Q, K, and V matrices, as shown in (10)-(12).

$$Q^{K\times H} = W_q^{K\times D}L \tag{10}$$
$$K^{K\times H} = W_k^{K\times D}L \tag{11}$$
$$V^{V\times H} = W_v^{V\times D}L \tag{12}$$

In which, W_q, W_k, and W_v are parameter matrices for linear mapping, and Q, K, and V are matrices composed of query vectors, key vectors, and value vectors, respectively.

(2) The dot product of the transpose matrices of Q and K produces the musical emotion feature similarity score matrix $Score^{H\times H}$. To address the issue of imbalanced softmax distributions resulting in small gradients when the dot product result is large, the dot product result is smoothed by scaling it with the square root $\sqrt{K}$ of the row-wise scaling of matrix Q, as shown in (13).

$$Score = \frac{QK^T}{\sqrt{K}} \tag{13}$$

(3) Softmax is applied to normalize the musical emotion similarity score matrix Score into the probability distribution matrix. The probability distribution matrix is then multiplied element-wise with matrix V to obtain the complete critical feature matrix $A^{V\times H}$ about music emotions, as shown in (14).

$$A = VSoftMax(Score) \tag{14}$$

C. LOSS FUNCTION

As an essential part of deep learning-based model training, loss functions such as Mean Squared Error (MSE) and Mean Absolute Error (MAE) are commonly used in regression problems. MAE is insensitive to outliers, and the gradient does not decrease with the decrease of the loss value during the gradient update process, which is not conducive to model convergence. On the contrary, MSE is more sensitive to outliers, and the gradient decreases as the loss value decreases during the gradient update process, which is beneficial to model convergence. Outliers refer to a very small portion of data with distribution patterns significantly different from the main data, often containing the trends of things. Therefore, outliers cannot be simply equated with noise [16]. Considering the complex and diverse features of music emotions, outliers in music emotion information may represent sudden changes in music emotions, but they may also be noise data. Considering the sensitivity to outliers and convergence, MSE is chosen as the loss function for model training in this paper. Its calculation is shown in (15):

$$MSE(i) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{15}$$

where N is the total number of musical emotion data points, y_i is the ground truth of the i-th musical emotion data point, and $\hat{y}_i$ is the regression value of the i-th music emotion data point.

IV. EXPERIMENTS

A. EXPERIMENT SETTINGS

The experiment is conducted based on the audio data of the EmoMusic dataset [17], the DEAM dataset [18] and PMEmo dataset [30]. To ensure that the musical emotion feature information analyzed by the proposed muSi-ABC model adheres to standardization, the eGeMAPS feature set, which has been validated and achieved significant results by researchers, was selected as the standard. The music emotional features were extracted from the audio data based on this feature set.

1) DATASET

The EmoMusic dataset, DEAM dataset, and PMEmo dataset were used in the experiment to train and evaluate the

effectiveness of the proposed muSi-ABC architecture in predicting musical emotions. The EmoMusic dataset consists of 744 songs, and 45-second music slices were extracted starting from 15 seconds of each song. The slices were annotated by Amazon Mechanical Turk workers, with at least 10 annotations per slice. Each slice was labeled with a static Valence-Arousal (VA) value and dynamic VA values at intervals of 0.5 seconds. The DEAM dataset expanded the EmoMusic dataset to 1744 songs. Besides the increased number of songs, the annotation mode and the length of music slices remained the same. To validate the generalizability of the proposed method, in addition to the two Perceived-style collected datasets mentioned above, the Induced-style collected PMEmo dataset was also used. The PMEmo dataset contains 794 full songs, and similar to the above two datasets, the annotation in the PMEmo dataset was done with the slider to collect dynamic annotations at a sampling rate of 2 Hz. Additionally, annotators should make a static annotation for the whole music excerpt on a nine-point scale after finishing dynamic labeling. To obtain long term musical emotion information, static musical emotions were predicted based on continuous time, and the ground truth labels were normalized to the [0, 1] range. In addition, the whole dataset was randomly divided into two parts, i.e., training set and testing set, in an 8:2 ratio.

2) FEATURE EXTRACTION

The eGeMAPS feature set was used as the standard for extracting music emotion features. This feature set is an audio emotion feature set that consists of 88 statistical acoustic features derived from 7 spectral features, 11 frequency-related features, and 7 energy/amplitude-related features through statistical calculations [22]. The features in the set and their correlations have been theoretically and practically validated, making it a standardized audio emotional feature set [23]. It is widely used in research related to audio emotion prediction [24] and music emotion prediction [25]. Based on the eGeMAPS feature set, the OpenSmile tool was used to extract continuous-time music emotion features from the audio dataset. The time term length, which is defined as the total length of different time sequences based on different frame intervals of the same song. Larger frame intervals generate long-term sample data, while smaller frame intervals generate long-term sample data. In this paper, a simplified variation approach was taken, considering only sample features of different term lengths for music emotion prediction, without considering the rationality of frame intervals, and ignoring the last frame information. Each song was represented in the form of time $\times$ features and saved in the .csv file. In addition, the advantages of feeding feature sets into the two-dimensional ConvNets are as follows:

1) Efficient feature extraction: Feature sets (such as eGeMAPS in this paper) provide pre-computed audio features that have been carefully selected and processed to capture key information from the audio. Using feature sets instead of raw audio signals reduces the computational requirements and number of parameters in the network, thus improving efficiency and training speed.

2) Enhanced representational power: Feature sets contain rich audio features that capture information related to different frequencies, temporal features, and semantics. By employing the two-dimensional ConvNets, we can leverage convolutional layers to model the spatial relationships of these features locally and globally, extracting more discriminative representations. This helps capture the structure, patterns, and crucial emotion-related information in the audio.

3) Network interpretability: Using feature sets as input makes it easier to understand and interpret the network's predictions. Since the features in the set have semantic interpretations, we can infer the network's attention to different audio attributes and emotion dimensions based on these features, thereby increasing the model's interpretability.

3) MODEL PARAMETERS AND EVALUATION METRICS

The experimental setup included using the Adam optimization algorithm, a weight decay coefficient of 0.0001, a learning rate of 0.0001, and a batch size of 4 samples. ReLU was used as the activation function in the model, and the number of training epochs was set to 80. Based on the eGeMAPS feature set, 88-dimensional features were extracted from the source music. Taking a time term of 99 as an example, the specific parameters of the model are shown in Table 1, where both the ConvNet and BiLSTM parts consist of two neural network layers. Connection was used to avoid repeated representation of input and output layers since the output of the previous layer serves as the input for the next layer. The SA module used $Q{=}K{=}V$, and the output layer's temporal dimension information was aggregated using the summation method. The batch size, which is the first dimension of each tensor and has the same value, is not presented in the model training parameters. Root Mean Square Error (RMSE) was used as the accuracy metric, and R2 (R-Squared) was used as the fitting metric. Additionally, we also used the concordance correlation coefficient (CCC), which focuses more on the dynamic trends of prediction results, as the metric.

TABLE 1. Parameter description of the model.

Model	Input / Output	Input[(height, width)] ->□Input[(99,88)]
ConvNet	Input	[(1, 99, 88)]
	Connection	[(4, 49, 44)]
	Output	[(4, 24, 22)]
BiLSTM	Input	[(24, 4×22)]
	Connection	[(24, 72)]
	Output	[(24, 32)]
SA	Input	[(24, 32)]
	$W_{q,k,v}$ Linear	(in = 32, out = 32)
	Output	[(32)]
Linear	Input	[(32)]
	Output	[(1)]

B. ABLATION EXPERIMENT

Due to the uncertainty of whether the proposed muSi-ABC architecture can improve the training efficiency and accuracy of music emotion prediction, experiments were conducted to verify the effectiveness of the muSi-ABC architecture and its components by extracting music emotion features with different temporal distance lengths from the EmoMusic dataset. The muSi-ABC architecture and its ablation models were tested. Firstly, the BiLSTM was used as the baseline model. Then, the two-dimensional ConvNet model and SA module were added to the BiLSTM separately. Finally, the ablation models obtained were BiLSTM (i.e., musi-B), ConvNet-BiLSTM (i.e., musi-BC), and BiLSTM-SA (i.e., muSi-AB). Ablation experiments were conducted to evaluate the prediction accuracy (i.e., RMSE), goodness of fit (i.e., R^2), and training efficiency on data with time term lengths of 99, 199, and 299. The RMSE and R^2 of each model with the minimum loss during training were compared, and the training efficiency (TE) was calculated as the ratio of the total training time to the total number of training epochs, representing the time required for one training epoch in seconds.

Based on the different term lengths, the regression evaluation results of the ablation models in the valence dimension and arousal dimension are shown in Table 2 and Table 3. The proposed muSi-ABC architecture outperformed the other three ablation models in terms of prediction performance on datasets with three different term lengths. Moreover, the prediction accuracy improved as the term length increased.

TABLE 2. The regression evaluation results of each ablation model in the valence dimension.

$Model_{Valance}$ /Term length		musi-B	musi-BC	muSi-AB	muSi-ABC
99	RMSE	0.0893	0.0894	0.0850	0.0850
	R^2	0.497	0.466	0.516	0.515
	TE/s	5.5	3.9	5.8	4.3
199	RMSE	0.0921	0.0880	0.0839	0.0832
	R^2	0.499	0.503	0.531	0.555
	TE/s	9.0	4.7	9.2	5.2
299	RMSE	0.0968	0.0862	0.0827	0.0825
	R^2	0.351	0.456	0.584	0.567
	TE/s	12.3	5.8	12.4	6.1

1) ANALYSIS OF THE EFFECTIVENESS OF THE TWO-DIMENSIONAL ConvNet AND SELF ATTENTION

Taking Arousal at the term length of 99 as an example, compared to using musi-B, musi-BC and muSi-AB reduced the RMSE by 0.0042 and 0.0098, respectively. The result from Table 3 indicates that the inclusion of the two-dimensional ConvNet and Self Attention has a positive effect on improving the prediction accuracy. In terms of the model performance of fusing SA across the three term lengths, BiLSTM-SA reduced the RMSE by 0.0098, 0.0093, and 0.0123, respectively, compared to BiLSTM. In terms of the model performance of fusing ConvNet across the three term lengths, ConvNet-BiLSTM reduced the RMSE by 0.0042, 0.0056, and 0.0089, respectively, compared to BiLSTM. Additionally, the training efficiency decreased by 1.7, 4, and 6.4 for ConvNet-BiLSTM across the three term lengths. These results further demonstrate the beneficial impact of combining Self Attention and the two-dimensional ConvNet in enhancing the overall performance of the final muSi-ABC architecture.

TABLE 3. The regression evaluation results of each ablation model in the arousal dimension.

$Model_{Arousal}$ /Term length		musi-B	musi-BC	muSi-AB	muSi-ABC
99	RMSE	0.0853	0.0811	0.0755	0.0755
	R^2	0.616	0.632	0.695	0.646
	TE/s	5.5	3.8	5.9	4.1
199	RMSE	0.0837	0.0781	0.0744	0.0744
	R^2	0.622	0.675	0.716	0.698
	TE/s	8.8	4.8	9.1	5.1
299	RMSE	0.0852	0.0763	0.0729	0.0725
	R^2	0.594	0.679	0.733	0.712
	TE/s	12.2	5.8	12.8	6.4

2) ANALYSIS OF RMSE AND R^2 CURVES

Combining Table 3 and taking Arousal at a term length of 99 as an example, the RMSE of the overall muSi-ABC architecture is reduced by 0.0098, 0.0056, and 0 compared to musi-B, musi-BC, and muSi-AB, respectively, as shown in Figure 6.

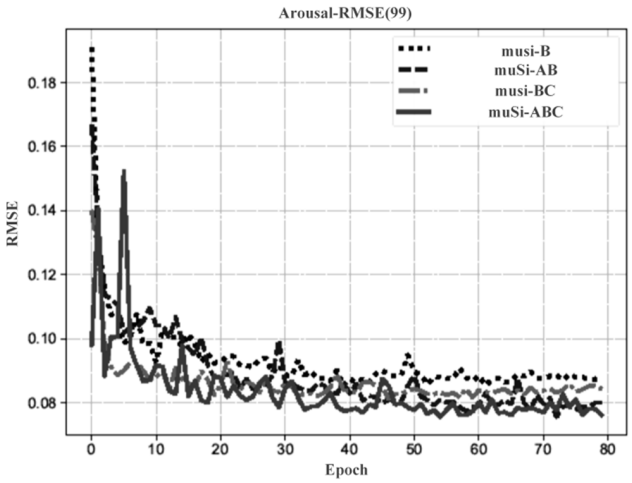

FIGURE 6. The RMSE curves for each model in Arousal with the term length of 99.

In Figure 6, although the final muSi-ABC architecture and muSi-AB architecture have the same RMSE under the minimum loss, the overall trend of the muSi-ABC architecture's RMSE is lower than that of muSi-AB. This result demonstrates that the overall prediction accuracy of the muSi-ABC architecture is higher than that of muSi-AB.

The R^2 of the muSi-ABC architecture relative to musi-B, musi-BC, and muSi-AB is increased by 0.03, 0.014, and -0.049, respectively, as shown in Figure 7.

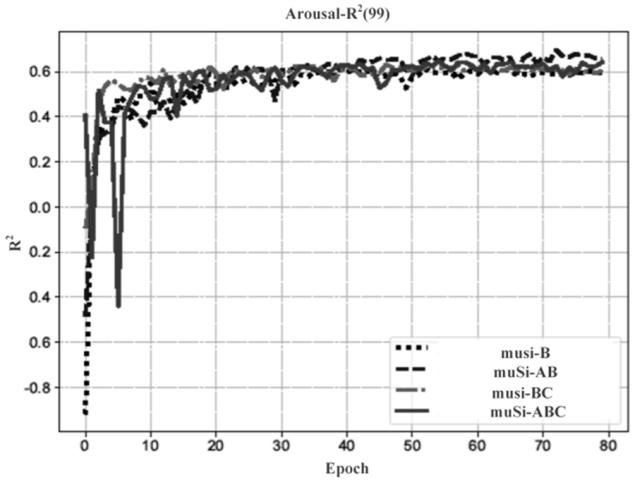

FIGURE 7. The R^2 curves for each model in Arousal with the term length of 99.

Although the final muSi-ABC architecture is 0.049 lower than muSi-AB, the R^2 is influenced by multiple factors. This result only indicates that the fitting performance of the proposed muSi-ABC architecture is slightly lower than that of muSi-AB and does not affect the comparison of their prediction accuracy.

3) ANALYSIS OF RMSE AND TRAINING EFFICIENCY FOR DIFFERENT TERM LENGTHS

Combining Table 3 and taking Arousal as an example, the variation of RMSE under the minimum loss for each model with increasing distance length is shown in Figure 8.

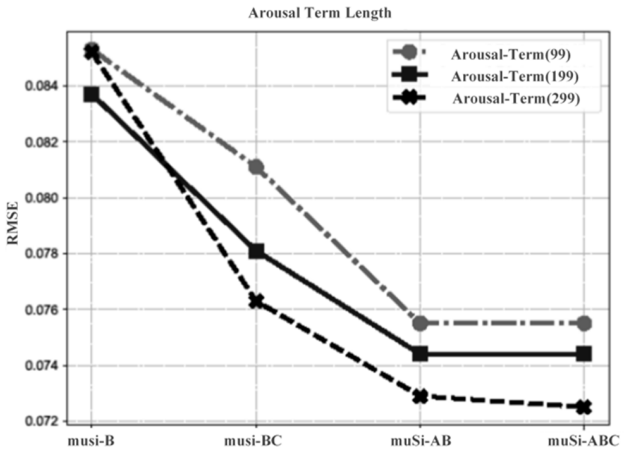

FIGURE 8. The RMSE for different term lengths in Arousal for each model.

As shown in Figure 8, for musi-B, the RMSE at a term length of 299 is 0.0015 higher than that at a term length of 199. This result indicates a decrease in learning ability for LSTM beyond a certain term length. Using muSi-AB, musi-BC, and the final muSi-ABC architecture, relative to a term length of 99, the RMSE values at term lengths of 199 and 299 are reduced by 0.0011, 0.0026, 0.003, 0.0048, 0.001, and 0.003, respectively. This result demonstrates that using long term data compared to short term data can improve prediction accuracy. As the term length increases, the prediction accuracy of the proposed muSi-ABC architecture gradually surpasses other ablated models, further confirming the muSi-ABC architecture's ability to improve the accuracy of long term musical emotion prediction.

In terms of training efficiency, Figure 9 clearly shows that the training efficiency of each model gradually increases with the term length.

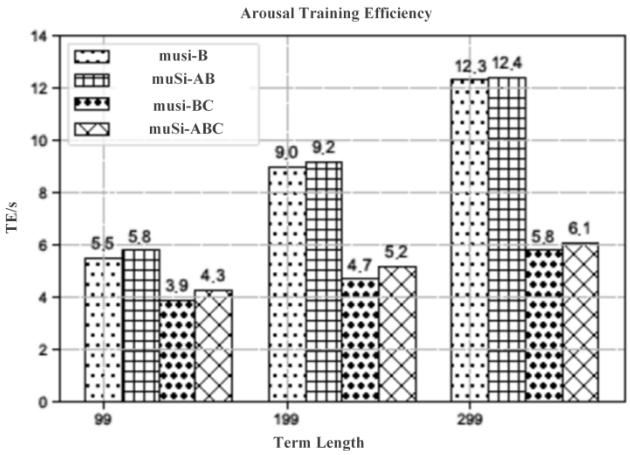

FIGURE 9. The training efficiency for different distance lengths in Arousal for each model.

Compared to muSi-AB, the muSi-ABC architecture's training efficiency is reduced by 1.8, 4, and 6.4 for different term lengths. This result demonstrates that integrating ConvNet can reduce model complexity and improve training efficiency.

In conclusion, although musi-BC has lower training efficiency than the muSi-ABC architecture, its RMSE is higher. muSi-AB has a similar RMSE to the muSi-ABC architecture, but lower training efficiency. Therefore, the construction of the muSi-ABC model, which simulates the perception process of music for people to express emotions, has certain advantages in predicting continuous-time, long term static music emotions, and can improve the accuracy and training efficiency of long term music emotion prediction.

4) COMPARISON OF PREDICTION ACCURACY BETWEEN BiLSTM AND ConvNet WITH DIFFERENT NUMBERS OF LAYERS

Adjusting the number of layers in BiLSTM and ConvNet networks to achieve higher music emotion prediction accuracy in the final muSi-ABC architecture. Firstly, experiments were conducted to determine the optimal number of layers in the BiLSTM network. Based on the determined number of BiLSTM layers, the number of ConvNet layers in the muSi-ABC model was determined. To ensure a suitable time term length, the RMSE at a term length of 199 was chosen as the evaluation metric for the model's layer configuration.

The RMSE of the BiLSTM network represents the prediction accuracy when using that network alone, while the RMSE of the 2D-ConvNet network represents the prediction accuracy when adjusting the number of ConvNet layers in the muSi-ABC architecture based on the determined number of BiLSTM layers.

The experiments compared the prediction accuracy of BiLSTM networks with 1 to 3 layers and the prediction accuracy of the muSi-ABC model using 1 to 3 layers of ConvNet. The goal was to identify the impact of the number of layers in BiLSTM and ConvNet on prediction accuracy. The experimental results are summarized in Table 4.

TABLE 4. Comparison of prediction accuracy for different BiLSTM and ConvNet layer numbers.

Model	Layer numbers	Valence RMSE	Arousal RMSE
BiLSTM	1	0.1013	0.0994
	2	0.0921	0.0837
	3	0.0961	0.0856
ConvNet	1	0.0866	0.0762
	2	0.0832	0.0744
	3	0.0895	0.0754

Taking Arousal as an example, the RMSE of the BiLSTM network with two layers was reduced by 0.0157 and 0.0019 compared to the network with one layer and three layers, respectively. Similarly, the RMSE of the ConvNet network with two layers was reduced by 0.0018 and 0.001 compared to the network with one layer and three layers, respectively. These results indicate that the prediction accuracy of the BiLSTM network with two layers and the ConvNet network with two layers is higher than the other configurations, and increasing the number of layers does not necessarily improve the prediction results. Therefore, based on the BiLSTM network with two layers and the ConvNet network with two layers, the ConvNet-BiLSTM model was constructed and combined with the self-attention model to form the muSi-ABC architecture for music emotion regression training.

5) THE IMPACT OF LOSS FUNCTIONS ON PREDICTING MUSICAL EMOTIONS

To verify whether the outliers in the music emotion features obtained from the eGeMAPS feature set are the turning points that affect the trend of music emotion changes, considering the complex and diverse nature of music emotion features and the sensitivity to outliers, MAE and MSE were used as the model training loss functions in the muSi-ABC architecture. RMSE was used as the evaluation metric for prediction accuracy. The experimental results are shown in Figure 10.

It is clear that using MSELoss for Valence and Arousal achieves good prediction accuracy compared to using MAELoss. Therefore, the music emotion features extracted based on the eGeMAPS feature set have standardization, and the outliers in the information contain the trend of music emotion changes, which can improve the model's prediction accuracy.

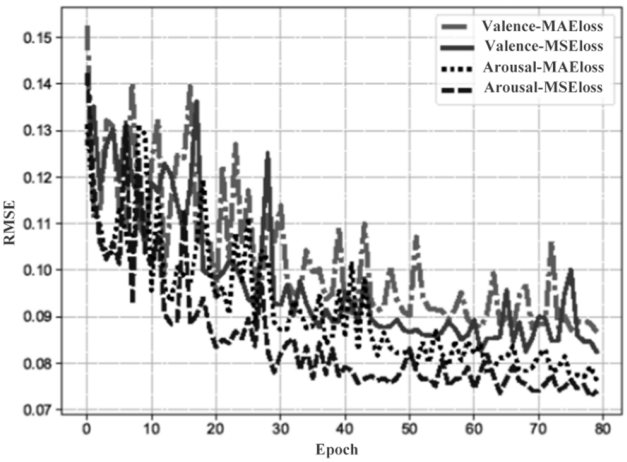

FIGURE 10. The impact of loss functions on predicting musical emotions.

C. CONTRAST EXPERIMENT

To further validate the performance effectiveness, the proposed method is compared with benchmark methods and state-of-the-art music emotion prediction methods using the EmoMusic dataset and the DEAM dataset, based on the same evaluation metrics. The following provides an overview of each comparative method:

MLR, BLSMT-RNN, SVR, and GPR [26]: These four models represent the benchmark prediction methods used for training and evaluating the EmoMusic dataset by the Technical University of Munich, Aizu University, and Utrecht University, respectively.

ConvNet_D-SVM [27]: This method explores the contextual information of emotion computation by increasing the receptive field of the network layers using dilated convolution (ConvNet_D) and feeding it into an SVM regression model.

AC2DConv [28]: This method analyzes audio features represented by a combination of raw audio, audio signals, and spectrograms using an audio and computed 2D convolution (AC2DConv) network model.

ResNets-audioLIME [29]: This method combines the source separation explainer audioLIME with residual networks (ResNets) to analyze intermediate perceptual features and spectrogram features.

The evaluation metric results of these methods are shown in Table 5.

Compared to other methods, the proposed method achieves the lowest RMSE and highest R^2 in music emotion prediction tasks. It improves the accuracy of music emotion prediction and exhibits the best fitting ability. Furthermore, to validate the generalization ability of the proposed method, comparative experiments were conducted on the PMEmo dataset, and the evaluation metric CCC was used. The results are also shown in Table 5. It can be seen that the fitting ability of the proposed method is still the best among the comparison

TABLE 5. Comparative experimental results of different methods on different datasets. Note that MLR* and SVR* are the baseline methods in the PMEmo dataset [30].

Datasets/Methods		Arousal RMSE	Arousal R^2	Valence RMSE	Valence R^2	Arousal CCC	Valence CCC
EmoMusic	MLR	0.12	0.48	0.15	0	0.501	0.474
	BLSMT-RNN	0.10	0.59	0.11	0.42	0.582	0.516
	SVR	0.10	0.63	0.12	0.35	0.603	0.584
	GPR	0.10	0.59	0.12	0.31	0.59	0.453
	ConvNet_D-SVM	0.10	0.63	0.11	0.41	0.612	0.509
	muSi-ABC	0.0775	0.7431	0.0857	0.5894	0.758	0.65
DEAM	AC2DConvStat	0.2003	0.5375	0.1928	0.162	0.568	0.51
	ResNets-audioLIME	0.25	0.51	0.21	0.54	0.551	0.483
	muSi-ABC	0.0815	0.6233	0.0791	0.5581	0.8	0.731
PMEmo	MLR*	0.096	0.457	0.143	0.122	0.694	0.587
	SVR*	0.097	0.511	0.129	0.144	0.753	0.664
	AC2DConvStat	0.087	0.677	0.093	0.322	0.701	0.574
	ResNets-audioLIME	0.083	0.791	0.090	0.389	0.658	0.551
	muSi-ABC	0.071	0.85	0.082	0.644	0.813	0.7

methods, and this result once again proves the effectiveness of the muSi-ABC model.

V. CONCLUSION

With the continuous advancement of music technology research, music emotions prediction has been widely applied in all kinds of fields. Inspired by this, we aim to simulate the perception process of music as people to express emotions. In response to the problems of long term dependencies and low training efficiency in music emotions prediction with LSTM neural networks, a novel and comprehensive network model called muSi-ABC is proposed for regression training of long-term music emotions prediction. Specifically, the proposed model uses the 2D-ConvNet to extract partial critical features of music emotions, employs the BiLSTM neural network to extract sequential information of music emotions from the obtained partial critical features, and utilizes the SA model to dynamically adjust the weights of the obtained sequential information, highlighting the complete critical points of music emotions. The ablation and contrast experimental results demonstrate that the proposed muSi-ABC model can reduce the training time for analyzing the regularities in music emotions information and effectively improve the accuracy of predicting music emotions. In conclusion, the proposed model for predicting musical emotions can capture the regularities of music emotions information from longer continuous durations, thereby improving prediction accuracy and training efficiency, and effectively achieving emotions music regression. It provides a new feasible idea for the direction of predicting music emotions.

The limitations of this study include the lack of consideration for additional modal information, and this may limit the generalization of the proposed method in capturing music emotions. In future research, we will explore the integration of audio data with listening data, lyrics text, and even video frames for a more comprehensive multimodal music emotion prediction.

REFERENCES

[1] S. R. Chiragkumar, "Chord recognition-music and audio information retrieval," 2021, *arXiv:2105.07019*.
[2] Z. Hu, L. Chen, Y. Luo, and J. Zhou, "EEG-based emotion recognition using convolutional recurrent neural network with multi-head self-attention," *Appl. Sci.*, vol. 12, no. 21, p. 11255, Nov. 2022.
[3] L. Jiang, H. Liu, H. Zhu, and G. Zhang, "Improved YOLO v5 with balanced feature pyramid and attention module for traffic sign detection," in *Proc. MATEC Web Conf.*, 2022, vol. 355, no. 6, p. 3023.
[4] L. Cai, S. Ferguson, H. Lu, and G. Fang, "Feature selection approaches for optimising music emotion recognition methods," 2022, *arXiv:2212.13369*.
[5] Y. Xiang, "Computer analysis and automatic recognition technology of music emotion," *Math. Problems Eng.*, vol. 2022, pp. 1–9, Mar. 2022.
[6] X. Yu, "Adaptability of simple classifier and active learning in music emotion recognition," in *Proc. 4th Int. Conf. Electron., Commun. Control Eng.*, Apr. 2021, pp. 13–19.
[7] C. Huang and Q. Zhang, "Research on music emotion recognition model of deep learning based on musical stage effect," *Sci. Program.*, vol. 2021, pp. 1–10, Oct. 2021.
[8] J. Dutta and D. Chanda, "Music emotion recognition in assamese songs using MFCC features and MLP classifier," in *Proc. Int. Conf. Intell. Technol. (CONIT)*, Jun. 2021, pp. 1–5.
[9] D. S. Naser and G. Saha, "Influence of music liking on EEG based emotion recognition," *Biomed. Signal Process. Control*, vol. 64, Feb. 2021, Art. no. 102251.
[10] W.-D. Dang, D.-M. Lv, R.-M. Li, L.-G. Rui, Z.-Y. Yang, C. Ma, and Z.-K. Gao, "Multilayer network-based CNN model for emotion recognition," *Int. J. Bifurcation Chaos*, vol. 32, no. 1, Jan. 2022, Art. no. 2250011.
[11] N. Satayarak and C. Benjangkaprasert, "On the study of Thai music emotion recognition based on western music model," *J. Phys., Conf. Ser.*, vol. 2261, no. 1, Jun. 2022, Art. no. 012018.
[12] Y. Liu Y, "Neural network technology in music emotion recognition," *Int. J. Frontiers Sociol.*, vol. 3, no. 1, pp. 1–10, 2021.
[13] F. Hasanzadeh, M. Annabestani, and S. Moghimi, "Continuous emotion recognition during music listening using EEG signals: A fuzzy parallel cascades model," *Appl. Soft Comput.*, vol. 101, Mar. 2021, Art. no. 107028.
[14] Z. Huang, S. Ji, Z. Hu, C. Cai, J. Luo, and X. Yang, "ADFF: Attention based deep feature fusion approach for music emotion recognition," 2022, *arXiv:2204.05649*.
[15] J. Qiu, C. L. Philip Chen, and T. Zhang, "A novel multi-task learning method for symbolic music emotion recognition," 2022, *arXiv:2201.05782*.
[16] K. Treerattanapitak and C. Jaruskulchai, "Outlier detection with possibilistic exponential fuzzy clustering," in *Proc. 8th Int. Conf. Fuzzy Syst. Knowl. Discovery (FSKD)*, vol. 1, Jul. 2011, pp. 453–457.
[17] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 songs for emotional analysis of music," in *Proc. 2nd ACM Int. Workshop Crowdsourcing Multimedia*, Oct. 2013, pp. 1–6.

[18] K. Sorussa, A. Choksuriwong, and M. Karnjanadecha, "Emotion classification system for digital music with a cascaded technique," *ECTI Trans. Comput. Inf. Technol.*, vol. 14, no. 1, pp. 53–66, Apr. 2020.
[19] P. Kantan, E. G. Spaich, and S. Dahl, "A technical framework for musical biofeedback in stroke rehabilitation," *IEEE Trans. Human-Mach. Syst.*, vol. 52, no. 2, pp. 220–231, Apr. 2022.
[20] B. Bahmei, E. Birmingham, and S. Arzanpour, "CNN-RNN and data augmentation using deep convolutional generative adversarial network for environmental sound classification," *IEEE Signal Process. Lett.*, vol. 29, pp. 682–686, 2022.
[21] K. Zhang, Z. Cao, and J. Wu, "Circular shift: An effective data augmentation method for convolutional neural network on image classification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 1676–1680.
[22] N. AbaeiKoupaei and H. Al Osman, "A multi-modal stacked ensemble model for bipolar disorder classification," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 236–244, Jan. 2023.
[23] W. Xue, C. Cucchiarini, R. W. N. M. van Hout, and H. Strik, "Acoustic correlates of speech intelligibility. The usability of the eGeMAPS feature set for atypical speech," Tech. Rep., 2019.
[24] H. Zhao, X. Zhou, and Y. Xiao, "Recognising continuous emotions in dialogues based on DISfluencies and non-verbal vocalisation features for a safer network environment," *Int. J. Comput. Sci. Eng.*, vol. 19, no. 2, pp. 169–178, 2019.
[25] D. Deutsch, L. Ray, M. Dolson, S. Zizook, and F. R. Moore, "Computer evaluation of musical performance from the acoustic signal: An exploratory study on performance anxiety," *J. Acoust. Soc. Amer.*, vol. 88, no. S1, p. S71, Nov. 1990.
[26] M. Soleymani, M. N. Caro, E. M. Schmidt, C. Sha, and Y. Yang, "1000 songs database," in *Proc. ACM Int. Workshop Crowdsourcing Multimedia*, 2014, pp. 4–7.
[27] A. Wadhawan and A. Aggarwal, "Towards emotion recognition in Hindi-English code-mixed data: A transformer based approach," 2021, *arXiv:2102.09943*.
[28] M. Sajid, M. Afzal, and M. Shoaib, "Multimodal emotion recognition using deep convolution and recurrent network," in *Proc. Int. Conf. Artif. Intell. (ICAI)*, Apr. 2021, pp. 128–133.
[29] S. Chowdhury, V. Praher, and G. Widmer, "Tracing back music emotion predictions to sound sources and intuitive perceptual qualities," 2021, *arXiv:2106.07787*.
[30] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Developing a benchmark for emotional analysis of music," *PLoS ONE*, vol. 12, no. 3, Mar. 2017, Art. no. e0173392.
[31] F. Korzeniowski, O. Nieto, M. McCallum, M. Won, S. Oramas, and E. Schmidt, "Mood classification using listening data," 2020, *arXiv:2010.11512*.
[32] Y. H. Yang and H. H. Chen, *Music Emotion Recognition*. Boca Raton, FL, USA: CRC Press, 2011.
[33] M. Schedl, E. Gómez, and J. Urbano, "Music information retrieval: Recent developments and applications," *Found. Trends Inf. Retr.*, vol. 8, nos. 2–3, pp. 127–261, 2014.
[34] L. Turchet and J. Pauwels, "Music emotion recognition: Intention of composers-performers versus perception of musicians, non-musicians, and listening machines," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 305–316, 2022.
[35] J. S. Gómez-Cañón, E. Cano, T. Eerola, P. Herrera, X. Hu, Y.-H. Yang, and E. Gómez, "Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications," *IEEE Signal Process. Mag.*, vol. 38, no. 6, pp. 106–114, Nov. 2021.

JING YANG was born in Zaozhuang, Shandong, in 1986. She received the Graduate degree from the Conservatory of Music, Shandong Normal University. Since 2011, she has been teaching with the Conservatory of Music, Qilu Normal University. Her research interests include cognitive computing and music cognition and music prediction. She is currently a member of Shandong Musicians Association.