Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **10**, Number **3**, August **2025**

Music Emotion Recognition Based on Feature Mapping and Convolutional Neural Networks

Yun Cao[1,*], Ying Liu[2]

[1]School of Music, Hohhot Minzu College, Hohhot 010000, P. R. China
papa3797@163.com

[2]College of Information Management, Mariano Marcos State University, Batac 2906, Philippines
wv6077@163.com

[*]Corresponding author: Yun Cao

Received July 8, 2024, revised November 26, 2024, accepted February 22, 2025.

Abstract. *Music emotions have complex diversity, and continuous dimensional emotion models map emotions to any point in the continuous dimensional space, with the ability to recognize rich and delicate emotions. The article conducts research on the V-A continuous dimension emotion model from two aspects: music emotion features and music emotion recognition models. Previous studies on music emotion recognition have concatenated features without verifying the correlation between features within the set and label truth values, which can increase model training time and reduce recognition accuracy. In response to this issue, this study divides music emotional features into two categories: low-level and mid to high-level. The extracted low-level music emotional features are fused with the filtered middle and high- level music emotional features to obtain a multi-level music emotional feature set. This article proposes a new fusion model CBSA (CNN BiLSTM Self Attention), which combines convolutional neural networks with attention mechanisms to effectively solve the problem of slow model training speed and improve the accuracy of music emotion recognition.*
Keywords: Music emotion recognition; Emotional characteristics of music; Two-dimensional convolutional neural network; Long Short-Term Memory Neural Network;

1. **Introduction.** Music, as the most common cultural and artistic form in human daily life, has permeated every corner of human life. Research on social development and human emotional needs has shown that it can directly describe human emotions and spiritual feelings [1]. The massive influx of music into people's lives provides a rich and diverse spiritual nourishment, but also increases the difficulty of obtaining music based on emotional information. In order to meet people's inner emotional needs in music, more and more researchers in related fields are paying attention to the research of quickly and effectively identifying music emotional information.

The process by which computers recognize music emotions through intelligent computing is called music emotion recognition [2]. It is hoped that through music emotion recognition models, the interconnection between music emotion features and music emotion models can be analyzed to describe music emotions, so that computers, like humans, have the ability to feel and understand music expression emotions. The mechanism of emotional generation originates from self-awareness [3]. In current research on the emotional process of music, human cognition is often used as the basis to divide the emotional characteristics of music into two categories: low-level and middle high-level. The low-level

features are not directly related to music emotions, but contain detailed information about music emotions; The characteristics of middle and high levels are directly related to music emotions, but ignore the details of music emotions. Meanwhile, most researchers in the field of Music Emotion Recognition (MER) extract music emotion features from audio files and directly input them into the recognition model. Although this ensures the richness and diversity of features, it does not take into account the degree of correlation between features and label truth values [4]. In summary, using only low-level or mid to high-level music emotional features has certain shortcomings. It is necessary to fully consider the diverse levels of music emotional features and select features that are highly correlated with music emotions, so that the recognition model can fully and comprehensively analyze music emotional features in order to achieve the best recognition effect.

1.1. **Related work.** In recent years, with the widespread application and in-depth exploration of human-computer emotional interaction technology, music emotion recognition research has gradually become a popular trend.

(1) Current status of research on music emotion recognition features

Music expresses emotions through basic elements such as pitch, length, strength, and timbre, and how to quantify them as musical emotional features is one of the key to solving the problem of music emotion recognition [5]. Based on human cognition, music emotional characteristics are divided into two categories: middle and high levels and low levels. There is a direct connection between the characteristics of middle and high levels and the emotional level of human cognition. Deng et al. [6] compared the accuracy of music emotion feature recognition between mid to high level and low level based on the same audio dataset. The experimental results showed that the accuracy of music emotion recognition using mid to high level features was improved by 10.2% compared to low level features, reaching 69.8%.

There is no direct connection between low-level features and human cognitive level emotions, but they can represent their detailed information and can be divided into low-level music emotional manual features and spectrogram features. Low level music emotional manual features often use manual methods to extract and select features, which requires a lot of manual and time costs in the screening and verification process. Doğdu et al. [7] outlined some audio emotion experts who have designed low-level audio emotion manual feature sets to reduce manual and time costs, such as the ComParE feature set, the 2009 Interspeech Challenge Set collection, and the eGeMaps feature set. At present, researchers often use the above feature sets for music emotion recognition research. Spectral features reflect the temporal changes of audio signals through spectral analysis of images, mainly applied in deep learning methods. They are input into recognition models, which can automatically analyze the emotional information contained in the features and obtain recognition results. Although this method reduces the cost of manually selecting features, it is prone to losing audio information during the information conversion process, has poor interpretability, and takes too long to train [8]. In response to this, Hizlisoy et al. [9] mainly used low-level music emotion manual features, supplemented by spectrogram features, and inputted them into a network model to recognize music emotions.

The extraction and selection of music emotional features is an indispensable part of music emotion recognition tasks. Choosing music emotional features that are highly correlated with label truth values can accurately and fully represent music emotions and improve overall recognition performance.

(2) Current research status of music emotion recognition methods

In related research, some scholars have used Support Vector Machines (SVM) [10, 11] or a combination of SVM and other statistical probability models for music emotion

classification training [12, 13, 14]. Although good recognition results have been achieved, there is uncertainty in sentiment classification standards. Yang et al. [15] addressed this issue by utilizing the characteristics of musical emotions and using regression training methods to improve recognition accuracy.

For recurrent neural networks, Grekow [16] proposed combining model tasks with LSTM and implemented the model using the WekaDeeplearning4j software package. The experiment proved the significance of using this method to identify emotions in music, and the results even exceeded those of SVM algorithms used for regression. Some scholars use convolutional neural networks. Considering the impact of local key information on music emotions, He and Ferguson [17] proposed incorporating convolutional neural networks into the model, which can reduce the complexity of manual annotation and improve the efficiency of feature mapping. Sams and Zahra [18] believe that when performing music classification tasks, a convolutional long short-term memory network can be added. By combining and superimposing the network, the extraction and recognition of music emotional features can be well completed.

A neural network that integrates attention models. Qiao et al. [19] created an EEG dataset consisting of four different types of music stimuli, and then constructed an LSTM based recognition model, which improved its recognition performance by utilizing the global perception ability of self attention mechanism. Zhang et al. [20] addressed the network problem in emotion recognition by using a composite attention network to reorganize it.

1.2. **Motivation and contribution.** In response to the problem of poor model fitting ability caused by the excessive use of music emotional features, this article uses feature mapping to screen music emotional features that are highly correlated with label truth values for optimization. To address the issue of insufficient and incomplete representation of music emotions using only low-level or mid to high-level features, multi-layer music emotion features are used for improvement. The article uses OpenSMILE tool to extract low-level music emotional features from audio files, and MIRToolBox tool to extract middle and high-level music emotional features. A CBSA model is constructed to address the problem of identifying the starting memory point of long-distance music emotion forgetting using long short-term memory neural networks, enabling computers to have the ability to recognize music emotions similar to humans. Therefore, the article analyzes multi-layer music emotional features based on the CBSA model to achieve recognition of continuous dimensional music emotions.

2. Analysis of relevant principles.

2.1. **Music emotional feature extraction tool.** The commonly used feature extraction tools in the field of Music Information Retrieval (MIR) include OpenSMILE, Librosa, MIRToolBox, PsySound, and Marsyas. Compared to other feature extraction tools, OpenSMILE can extract features contained in audio sentiment feature sets, while MIRToolBox can extract high-level semantic features of music in audio data.

The OpenSMILE [21] tool uses an operation command similar to "SMILExtract - C% configfile - I% inputfile - csvoutput% output.csv" to extract features based on relevant feature configuration files, where % configfile is the feature configuration file, % inputfile is the input audio file, and csvoutput table outputs the extracted feature data in csv format. OpenSMILE is suitable for audio related fields such as speech recognition, music emotion recognition, and music information retrieval. It can perform four types of operations: data signal processing, data preprocessing, and extracting low-level manual features of audio and video.

The MIRToolBox tool [22] is a Matlab toolbox developed by interdisciplinary researchers in music, neuroscience, emotional computing, and other fields, which extracts high-level semantic features of music such as timbre, pitch, beat, and rhythm from audio files. The MIRToolBox tool manual categorizes music features into five categories based on their constituent elements: Dynamics, Rhythm, Timber, Pitch, and Tonality, including approximately 50 audio, music feature extraction functions, and statistical descriptors.

2.2. **Support vector regression.** Support Vector Machine maps data from nonlinear space to linear feature space through kernel function, and searches for the best classification surface, i.e. hyperplane, in this space to achieve sample classification. Support vector regression is the application of SVM to regression problems. SVR has good decision-making ability on data and can be applied to the study of continuous dimensional music emotional features.

Suppose there is a music emotion dataset $D = \{(x_1, y_1), ..., (x_n, y_n)\}$, where x_i represents the music emotion feature vector of the i-th song, y_i represents the label truth value corresponding to x_i, and n is the total number of songs in the dataset. The linear regression discriminant model is represented as $f(x)$, with the ultimate goal of making it as close to the true value as possible. The calculation of $f(x)$ is shown in Equation (1).

$$f(x) = w^T x + b \tag{1}$$

In the research and calculation process, it is necessary to keep the distance between the hyperplane and the farthest sample point as small as possible, that is, when the loss function L_ε is the smallest, the model fits best. The specific calculation is shown below.

$$m = f(x_i) - y_i \tag{2}$$

$$L_\varepsilon(m) = \begin{cases} 0, & \text{if } |m| < \varepsilon \\ |m| - \varepsilon, & \text{otherwise} \end{cases} \tag{3}$$

where m represents the error between the predicted value $f(x_i)$ and the true value y_i, and as long as the predicted value falls within the $|m| < \varepsilon$ region, no loss is considered.

When the model fits best, the objective function is minimized, and $f(w, b)$ is the desired value. The calculation formula is as follows.

$$\min f(w, b) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} L_\varepsilon(f(x_i) - y_i) \tag{4}$$

where C is the penalty term, indicating the tolerance for error m, and the larger C, the greater the tolerance for m. When $f(x_i) = y_i$, x is located in the hyperplane of the linear decision boundary.

In practical problems, most data samples are nonlinear and indivisible, and kernel functions can be used to solve this problem. The specific calculation is shown in Equation (5), where x is the feature vector in the original space, x' is the new feature vector mapped to a high-dimensional space, and σ is the Gaussian kernel bandwidth. Equation (6) is the SVR discriminant model with added kernel function.

$$k(x, x') = e^{-\frac{||x-x'||^2}{2\sigma^2}} \tag{5}$$

$$f(x) = k(x, x') + b \tag{6}$$

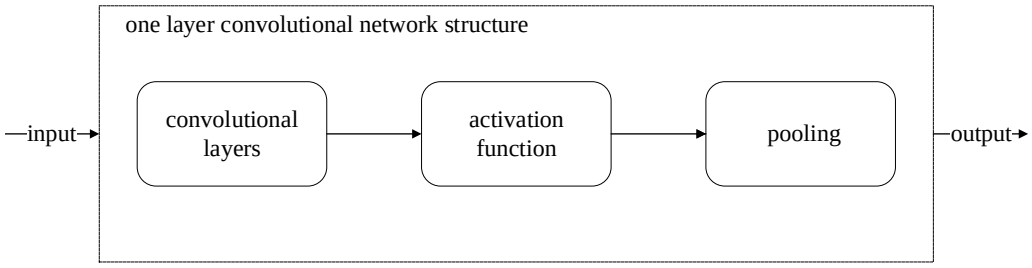

Figure 1. One layer convolutional network structure

2.3. **Convolutional neural network.** Convolutional neural networks can extract local features with rich deep meanings from raw data.

The internal process is to use convolutional kernels to perform convolution operations on input features, use activation functions to perform non-linear feature mapping on the convolution operation results, and use pooling layers to perform pooling operations on the aforementioned non-linear features.

(1) Convolutional layer

Taking a two-dimensional convolutional neural network as an example, assuming the input two-dimensional feature data i, the convolution kernel is w, and the convolution definition is shown in Equation (7). Among them, b is a scalar bias, and o is the two-dimensional feature data obtained after convolution operation.

$$o_{x,y} = \sum_{m=0}^{M} \sum_{n=0}^{N} (w_{m,n} i_{x+m,y+n} + b) \tag{7}$$

(2) Activation function

The activation function maps the input linear neurons of the network model to the output through nonlinear relationships, enhancing the neural network's ability to learn and simulate complex data features, and enhancing its fitting ability. The commonly used activation functions include Tanh, Sigmoid, and ReLU, among which ReLU is currently the commonly used activation function in deep learning. The definition of ReLU is as follows: when $x > 0$, the derivative is 1, which to some extent alleviates the problem of gradient vanishing in neural networks and improves convergence speed.

$$\mathrm{ReLU}(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{8}$$

Input the convolution result of Equation (7) into the nonlinear activation function $f(\cdot)$ to obtain the output feature $p_{x,y}$. The specific calculation is as follows.

$$p_{x,y} = f(o_{x,y}) \tag{9}$$

The pooling layer utilizes the idea of local correlation to perform pooling operations on a set of locally correlated features, obtaining new feature data. On the basis of retaining local key features, reducing feature dimensions and data space size, thereby reducing the number of network model parameters, computational complexity, and overfitting risks, and accelerating calculation speed. There are two commonly used pooling operations: Mean Pooling and Max Pooling. Mean pooling is the process of selecting the average sum of all feature points within the pooling window, which is suitable for preserving all information; Maximum pooling is the process of selecting the maximum value of all feature points within the pooling window, which is suitable for retaining certain key information.

2.4. **Attention mechanism.** Currently, the use of attention mechanism is almost indispensable in each model, it was first proposed in the field of computer vision, and then the Google mind team combined the attention mechanism and RNN on the task of image categorization, so that the attention mechanism began to develop rapidly.

3. Music feature mapping and CBSA model.

3.1. Music emotional feature mapping.

3.1.1. *Extracting emotional features from low-level music.* In response to the problem of a large number of features and insufficient flexibility of the recognition model, which leads to low recognition accuracy, the eGeMAPS was selected as the dataset. Common music emotional features and the content of the eGeMAPS feature set were introduced in the relevant theoretical knowledge. Compared to feature sets with more features, the eGeMAPS feature set has fewer features, making it easier for model training; Compared to feature sets with fewer features, feature representation is more complete. Using a professional audio emotion feature set can avoid the problems of low effectiveness and high resource cost consumption in customizing low-level manual features.

The GeMAPS feature set consists of a total of 62 features, all of which are HSF features. HSF features are statistical features obtained through processing based on manually set features, such as mean and maximum values. The music dataset used in this article is an extension of GeMAPS, which adds some features on top of 18 artificial features, including 5 spectral features: MFCC1-4 and Spectral flux (spectral difference between two adjacent frames) and 2 frequency related features: the bandwidth of the second and third resonance peaks. After performing arithmetic mean and standard deviation normalization on these 7 features, 14 feature values can be obtained. Then, the relevant feature values of the silent zone are calculated, along with the feature values of the equivalent sound level, resulting in a total of 88 features.

3.1.2. *Emotional feature extraction in mid to high level music.* The emotional characteristics of middle and high-level music are closer to emotions and have richer information content. In terms of feature extraction, it is more complex to manually calculate middle and high-level features from WAV files, so the MIRToolBox tool is chosen to extract features. Under the guidance of professional musicians, based on the knowledge of emotional features in mid to high level music introduced in relevant theories, combined with the emotional features of mid to high level music that can be extracted using the MIRToolBox tool, eight dimensional features of pitch, tempo, pulse integrity, key, brightness, roughness, regularity, and music state change are extracted from audio data.

3.2. **Selection of emotional features in music.** Although computer technology is constantly advancing and neural networks are constantly evolving, the flexibility of models still lags far behind the human brain's ability to respond to things. To improve recognition accuracy, in addition to enhancing the recognition ability of the network model, it is also necessary to streamline features that are weakly related to the current task, thereby reducing the computational cost of the model. This method of simplifying useless features is feature selection.

The feature selection method includes three methods: filtering, wrapping, and embedding. The filtering method is based on statistical theory and scores the correlation between features and label truth values from the perspective of features. The calculation is simple and independent of the model, only considering the correlation between a single feature and label truth values; The wrapping method selects different feature subsets from the initial feature set, trains the learner, and selects the optimal feature

subset based on the learner's evaluation indicators, which is costly and only considers the correlation between feature combinations and label truth values; The embedding method takes feature selection as a part of model training, which involves first using the model for training, obtaining various feature weights, and selecting features based on the size of the weights. In the method, except for features with weights of 0, which are not useful for recognition accuracy, other features are difficult to define, so the process of selecting and evaluating the model can be very time-consuming and laborious. The eGeMAPS feature set is a low-level audio sentiment feature set with standard normativity, therefore feature selection is not performed on it. Only feature selection is performed on the feature combinations extracted by the MIRToolBox tool. In order to obtain the optimal feature subset and reduce the time and labor costs of screening and verification, considering the small scale of feature combinations, from the perspective of the correlation between the combined features and the true value of individual features on the label, the wrapping method and filtering method are used for feature selection. During the screening process, time and data augmentation factors are not considered. The emotional feature matrix of each song mentioned above is averaged based on the time dimension and transformed into a feature vector form. The feature vector of the i-th song is defined as where $1 \leq p \leq 10$. $g_{i,p} = [g_{i,1}, \ldots, g_{i,p}]$

3.3. **Fusion of emotional features in music.** Feature fusion is the operation of generating a new set of features by fusing multiple features with different characteristics. From the perspective of fusion processing time, feature fusion can be divided into early fusion and late fusion. Early fusion performs multi-layer feature fusion before the features are input to the recognition model; Late fusion fuses the recognition results of different features at the decision-making level. Early fusion includes series feature fusion and parallel strategy fusion. The former connects multiple features in a concatenated form, while the latter combines two feature vectors into a complex vector form. In order to obtain multi-level music emotional features while keeping the time dimension unchanged, the series feature fusion method in early fusion is used to concatenate the low-level and high-level features of music emotions, resulting in multi-layer music emotional features. The specific description is as follows:

(1) After feature selection, a new feature combination matrix $N_{X\times P} = \{n_1, \ldots, n_X\}$ is obtained from the extracted feature combination matrix $N_{X\times M} = \{n_1, \ldots, n_X\}$, where M is the feature dimension.

(2) Assuming the function of the series of feature fusion methods is concat(), the low-level music emotion feature matrix $H_{X\times Y}$ and $N_{X\times M}$ extracted in Section 3.3.1 are fused, with the time dimension unchanged and the feature dimension changed to $N = M + Y$, to generate multi-layer music emotion feature below. $I_{X\times N} = \{i_1, \ldots, i_X\}$ The expression for feature fusion is shown:

$$I_{X\times N} = \text{concat}(H_{X\times Y}, N_{X\times M}) \tag{10}$$

3.4. **CBSA model construction method.** This article represents each song as an $I_{X\times N}$ music emotional feature matrix, where X represents the time and N represents the music emotional feature. The overall structure of the model is shown in Figure 2.

The CBSA model simulates the specific process of human perception of music expression emotions. It uses a two-dimensional CNN to obtain the main melody fragments in the song, and then uses BiLSTM to calculate the relevant features of music emotions through forward propagation of the network. Then, the Self Attention (SA) model combines the above information to obtain global key music emotion information.

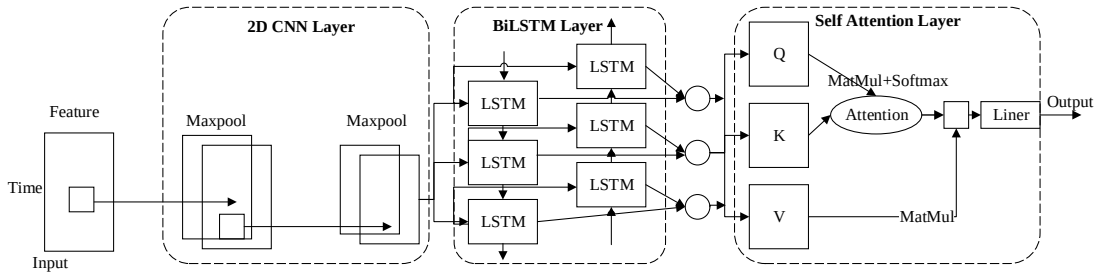

Figure 2. The overall structure of the CBSA model

Table 1. Formal definition based on each output layer of a song

Formal definition description	**Formal definition**
The input form for each song is a matrix I of $X \times N$	$I_{X\times N} = \{i_1, ..., i_X\}$
The output of a 2D convolutional layer is a matrix N of $A \times B$	$N_{A\times B} = \{n_1, ..., n_A\}$
BiLSTM outputs a matrix L of $D \times H$	$L_{D\times H} = \{l_1, ..., l_D\}$
The output of the self-attention layer is a matrix A of $V \times H$	$A_{V\times H} = \{a_1, ..., a_V\}$

3.5. **2D Convolutional Layer.** Two dimensional convolutional layers have two dimensions: height and width. In this article, we correspond them to time and feature, respectively. Here is an example to further illustrate that in the input stage, the sample matrix is input into the model network, and after mutual calculation with the kernel array, the feature matrix is obtained. The results are then obtained through normalization, activation function, and pooling operations. At this point, this result is not the final result, it is only the result obtained after passing through a layer of network. Then, repeat the above operation process again, and the final result obtained is the feature matrix we need.

3.6. **BiLSTM.** Assuming the three "gates" and two states of the LSTM loop unit structure, namely the forget gate g_t, input gate k_t, output gate p_t, internal state c_t, and candidate state c'_t, the external state at time t is j_t, and the external state at the previous time is j_{t-1}. The LSTM calculation process is as follows:

$$g_t = \sigma(E_g n_t + I_g j_{t-1} + b_g) \tag{11}$$
$$k_t = \sigma(E_k n_t + I_k j_{t-1} + b_k) \tag{12}$$
$$p_t = \sigma(E_p n_t + I_p j_{t-1} + b_p) \tag{13}$$
$$c'_t = \tanh(E_c n_t + I_c j_{t-1} + b_c) \tag{14}$$
$$c_t = k_t \cdot c'_t + g_t \cdot c_{t-1} \tag{15}$$
$$j_t = p_t \cdot \tanh(c_t) \tag{16}$$

where E_x is the current time weight matrix, I_x is weight matrix, b_x is the bias vector, σ is the sigmoid function.

On the basis of unidirectional LSTM, Equations (17) and (18) are used to extract and store past and future music sentiment information, respectively, to achieve music sentiment data modeling based on BiLSTM. Assuming the current moment is t, the hidden layer states of the Forward and Backward layers are defined as h_1 and h_2. The output l_t of the BiLSTM network layer is calculated as:

$$h_1 = f(U_1 h_{t-1} + W_1 n_t + b_1) \tag{17}$$
$$h_2 = f(U_2 h_{t-1} + W_2 n_t + b_2) \tag{18}$$
$$l_t = W_{t1} h_1 + W_{t2} h_2 + b_0 \tag{19}$$

where W_x is the current time weight matrix, U_1 is weight matrix, U_2 is the next time weight matrix, W_{tx} is the current time hidden layer state weight matrix, b_x is the bias vector, and $f(\cdot)$ is the activation function used by the hidden layer.

3.7. **Self attention layer.** Take the emotional feature vectors of music at each moment in L as query vectors, score the similarity with the emotional feature vectors at different moments in music, and obtain the global key feature information of music emotions through weighted averaging.

(a) Perform linear mapping, and the specific calculation is as follows.

$$Q = E_Q L \tag{20}$$

$$K = E_K L \tag{21}$$

$$V = E_V L \tag{22}$$

where E_Q, E_K, and E_V are linear mapping value, respectively.

(b) The transpose matrix of K and the Q point are multiplied to obtain the Score matrix of music emotional similarity:

$$\text{Score} = K^T Q \tag{23}$$

(c) Normalize the Score matrix of music sentiment similarity using the SoftMax function:

$$A = V \cdot \text{SoftMax}(\text{Score}) \tag{24}$$

4. Experiment.

4.1. **Data set.** Select a music emotion dataset containing audio for the purpose of studying music emotion features; Select audio clips with longer time intervals for long-distance music emotion recognition research. Based on the above two points, the article selects the EmoMusic dataset, which contains the original audio data and has the longest audio segment, as the experimental basic data from the publicly available continuous dimensional music emotion dataset. The DEAM dataset is used to further validate the method proposed in the article. In order to obtain long-distance music emotional information, the experiment is based on continuous time recognition of static music emotions; To reduce the burden of model recognition, the label truth values are normalized to be distributed within the [0,1] interval.

4.2. **Experimental configuration.** Set the penalty term C for the experimental parameters of the SVR model to 1, with a random seed count of 66, and select a Gaussian kernel function as the kernel function. The learning rate of the CBSA model experiment is 0.001, with 4 samples per batch. After feature selection, there are 6 remaining mid to high level features, and the fused features are 94 dimensions. Taking the 99-time distance as an example, the parameters of each layer in constructing the CBSA model are shown in Table 2.

4.3. **Evaluating indicator.** The experiment uses Root Mean Square Error (RMSE) as the accuracy indicator for recognition. The smaller the RMSE value, the smaller the deviation between the label true value and the regression value, and the higher the recognition accuracy. The specific calculation is shown below.

$$\text{RMSE}(i) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{25}$$

Table 2. Model parameter

Model Layer	**Input/Output**	**Input[(height,width)]**
Input		[(1, 99, 94)]
CNN	Connection	[(4, 49, 47)]
	Output	[(4, 24, 23)]
	Input	[(24, 4×23)]
Bi-LSTM	Connection	[(24, 78)]
	Output	[(24, 32)]
SA	Input	[(24, 32)]
	Linear (in = 32, out = 32)	
	Output	[(32)]
Liner	Input	[(32)]
	Output	[(1)]

Use R-Squared (R^2) as the goodness of fit indicator for the model. The larger the R^2 coefficient, the better the fit of the model, and the more data points fall within the regression line. The specific calculation is shown below.

$$R^2(i) = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2} \tag{26}$$

In equations above, N is the total amount of all data in the dataset, y_i is the true value, $\hat{y}_i$ is the predicted value, and $\overline{y}$ is the average of y.

4.4. **Results and analysis of CBSA model ablation experiments.** This section will conduct ablation experiments on the CBSA model proposed in this article, mainly targeting music emotional features of different distance lengths. The recognition tasks with different DL (distance lengths) pose challenges to the performance of the model, and the longer the distance, the higher the difficulty of music recognition. In this experiment, BiLSTM, CNN-BiLSTM, and BiLSTM-SA were used as baselines.

Taking Valence as an example, as shown in Table 3, the RMSE in Valence varies with the increase of DL, as shown in Figure 3. The RMSE of BiLSTM at distance length 299 increases by 0.0087 and 0.0055 compared to distance lengths 99 and 199. This result proves that LSTM has a decline in learning ability after exceeding a certain distance length. The other three models achieved better results over long distances, which proves that using long-distance data can improve recognition accuracy compared to short distances. As the distance length increases, the recognition accuracy of the CBSA model gradually exceeds that of other ablation models, this further proves that the CBSA model still has good performance in more complex scenarios.

In addition to comparing the RMSE of the models, we will also compare the learning speed. As shown in Figure 4, the blue part represents BiLSTM, the green part represents BiLSTM-SA, and the yellow part represents CNN-BiLSTM, all of which are the baselines used in the experiment. The red part represents the CBSA model proposed in this paper. Firstly, by comparing the blue and yellow parts, it can be seen that the learning speed of the yellow part has been significantly improved. Secondly, by observing the green and red parts, it can be seen that the learning speed of the red part has been improved. Through the above comparative experiments, it can be seen that adding CNN to the model can improve learning speed.

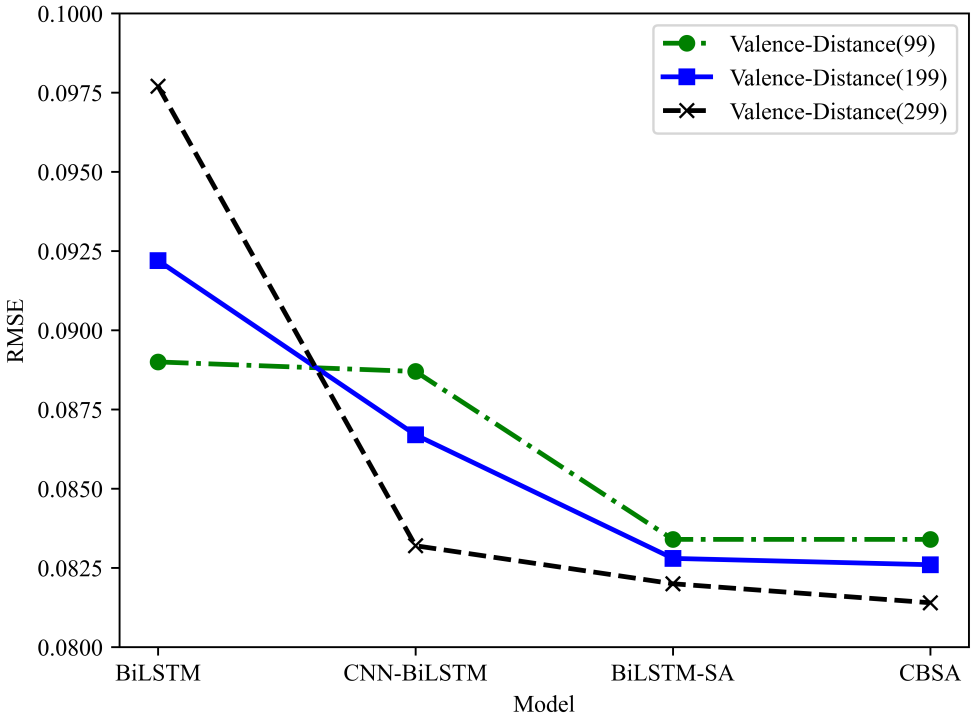

Figure 3. RMSE of different distance lengths in Valence models

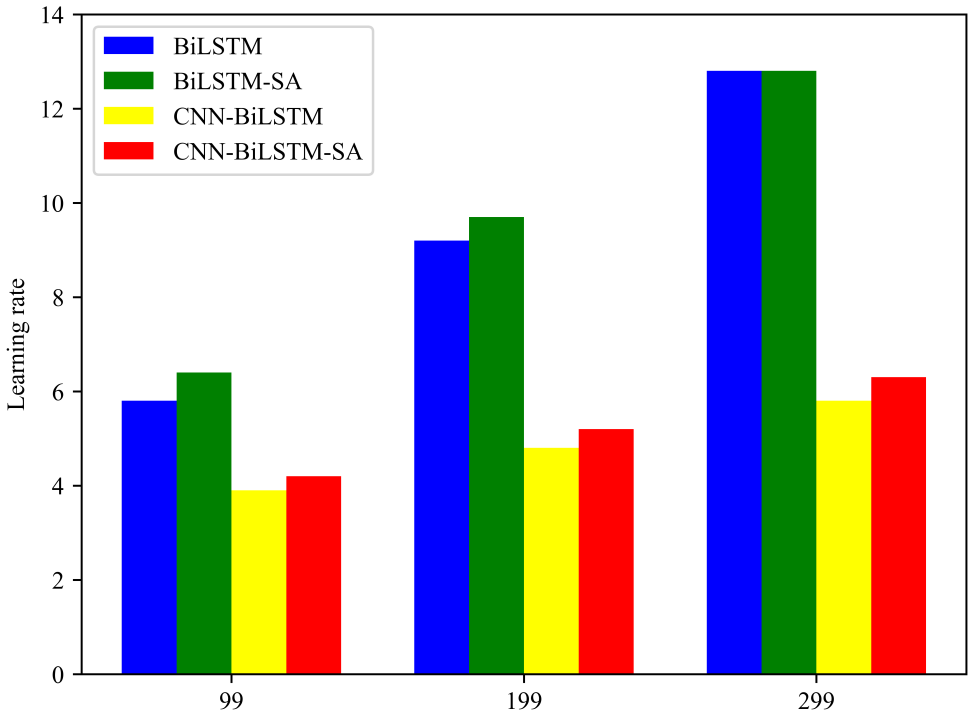

Figure 4. Training efficiency of different distance lengths in Valence models

4.5. **Comparative experimental results and analysis.** In recent years, there has been a lot of research on music emotion recognition. To further verify the effectiveness of the model performance, based on the same evaluation index, the method proposed in this paper is compared with the baseline and current methods with better performance. The following text provides a brief introduction to each comparison method.

(a) MLR [23] is the baseline method for the dataset used in this article.
(b) BLSMT-RNN [24] is a recognition method used in the training evaluation of the EmoMusic dataset at the Technical University of Munich.
(c) CBSA_eGeMAPS [25]. Based on the low-level audio emotion feature set of eGeMAPS, the CBSA model is used to recognize music emotions.
(d) CBSA. Using the CBSA model to analyze multi-level musical emotional features and achieve recognition of continuous dimensional musical emotions.

The evaluation index results are listed in Tables 3 and 4. According to Table 3, based on the EmoMusic dataset, taking the Valence dimension RMSE recognition results as an example, compared to the MLR, BLSMT-RNN, and CBSA_eGeMAPS methods, the recognition accuracy of the proposed method decreased by 0.0686, 0.0286, and 0.0011. This result proves that the proposed method has better recognition performance on the EmoMusic dataset. The R^2 of the method in the article is 0.02 lower than that of CBSA_eGeMAPS. This conclusion does not affect the comparison of recognition accuracy between the two methods, as R^2 is influenced by various factors such as features and datasets.

According to Table 4, based on the DEAM dataset, taking the RMSE recognition results of the valence dimension as an example, compared with the AC2DconvStat, ResNets audioLIME, and CBSA_eGeMAPS methods, this method shows significant improvement in performance. By comparing the data of R^2 in the table, it can be seen that the evaluation index results of R^2 are easily influenced by features and dataset factors, further indicating that using multi-level music emotional features can improve recognition accuracy.

Table 3. Summary of model accuracy based on the EmoMusic dataset

Dataset	Model	Arousal		Valence	
		RMSE	R^2	RMSE	R^2
EmoMusic	MLR	0.12	0.48	0.15	0
	BLSMT-RNN	0.10	0.59	0.11	0.42
	CBSA_eGeMAPS	0.0725	0.712	0.0825	0.567
	CBSA	0.0711	0.707	0.0814	0.547

Table 4. Summary of model accuracy based on DEAM dataset

Dataset	Model	Arousal		Valence	
		RMSE	R^2	RMSE	R^2
DEAM Dataset	AC2DConvStat	0.2003	0.5375	0.1928	0.162
	ResNets-audioLIME	0.25	0.51	0.21	0.54
	CBSA_eGeMAPS	0.0795	0.615	0.0784	0.551
	CBSA	0.0773	0.621	0.0767	0.555

5. **Conclusions.** Emotion recognition task is an important part of implementing artificial intelligence technology, and as part of its research, music emotion recognition is associated to the technologies such as music psychotherapy, music education, and music recommendation, which can promote the development of machine understanding of emotions. This study aims to identify rich and delicate music emotions, based on a continuous dimension emotion model, and conducts research on MER from both music emotion features and recognition models. The specific tasks are as follows:

This article provides an overview of the current research status and cutting-edge technological developments in music emotion both domestically and internationally, and briefly introduces the theory and technology of music emotion recognition involved in the article. Afterwards, research will be conducted on music feature mapping, extracting low-level music emotional features, and using the MIRToolBox tool to extract mid to high-level music emotional features based on music formation elements. Afterwards, we established a CBSA model and conducted experiments, which verified the correctness of our research approach.

The development of music emotion recognition technology has achieved certain research results, and many researchers have provided new ideas and methods for music emotion recognition, but it is still in the exploratory stage. Although the article has achieved some research results on continuous dimension music emotion recognition to a certain extent, due to the limitations of level and conditions, there is still a lot of research space in this article. The following issues can be further studied in the future.

The basic data used in this study is audio data, but some songs give people a feeling of happiness based on audio judgment, and after analyzing the lyrics, it is difficult to define the emotions of the songs. For this purpose, future research can consider a multimodal approach that combines audio, lyrical text, and video images to comprehensively describe music emotions from multiple perspectives and improve the accuracy of music emotion recognition.

REFERENCES

[1] L. Perlovsky, "Musical emotions: Functions, origins, evolution," *Physics of Life Reviews*, vol. 7, no. 1, pp. 2–27, 2010.

[2] D. Han, Y. Kong, J. Han, and G. Wang, "A survey of music emotion recognition," *Frontiers of Computer Science*, vol. 16, no. 6, 166335, 2022.

[3] X. Yang, Y. Dong, and J. Li, "Review of data features-based music emotion recognition methods," *Multimedia Systems*, vol. 24, pp. 365–389, 2018.

[4] R. Panda, R. Malheiro, and R. P. Paiva, "Novel audio features for music emotion recognition," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 614–626, 2018.

[5] R. Panda, R. Malheiro, and R. P. Paiva, "Audio features for music emotion recognition: a survey," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 68–88, 2020.

[6] Y. Deng, Y. Lu, M. Liu, Y. Cui, and Q. Lu, "Music emotion recognition model based on middle and high-level features," *Computer Engineering and Design*, vol. 38, no. 04, pp. 1029–1034, 2017.

[7] C. Doğdu, T. Kessler, D. Schneider, M. Shadaydeh, and S. R. Schweinberger, "A comparison of machine learning algorithms and feature sets for automatic vocal emotion recognition in speech," *Sensors*, vol. 22, no. 19, 7561, 2022.

[8] L. Gao, K. Xu, H. Wang, and Y. Peng, "Multi-representation knowledge distillation for audio classification," *Multimedia Tools and Applications*, vol. 81, no. 4, pp. 5089–5112, 2022.

[9] S. Hizlisoy, S. Yildirim, and Z. Tufekci, "Music emotion recognition using convolutional long short term memory deep neural networks," *Engineering Science and Technology, An International Journal*, vol. 24, no. 3, pp. 760–767, 2021.

[10] Y. Xia, L. Wang, and K.-F. Wong, "Sentiment vector space model for lyric-based song sentiment classification," *International Journal of Computer Processing of Languages*, vol. 21, no. 04, pp. 309–330, 2008.

[11] A. Krishnaiah, and P. B. Divakarachari, "Automatic Music Mood Classification using Multi-class Support Vector Machine based on Hybrid Spectral Features," *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 5, 2021.

[12] K. Sorussa, A. Choksuriwong, and M. Karnjanadecha, "Emotion classification system for digital music with a cascaded technique," *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, vol. 14, no. 1, pp. 53–66, 2020.

[13] D. Chaudhary, N. P. Singh, and S. Singh, "A survey on autonomous techniques for music classification based on human emotions recognition," *International Journal of Computing and Digital Systems*, vol. 9, no. 03, 2020.

[14] Y. Liu, "Neural network technology in music emotion recognition," *International Journal of Frontiers in Sociology*, vol. 3, no. 1, 2021.
[15] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 448–457, 2008.
[16] J. Grekow, "Music emotion recognition using recurrent neural networks and pretrained models," *Journal of Intelligent Information Systems*, vol. 57, no. 3, pp. 531–546, 2021.
[17] N. He, and S. Ferguson, "Music emotion recognition based on segment-level two-stage learning," *International Journal of Multimedia Information Retrieval*, vol. 11, no. 3, pp. 383–394, 2022.
[18] A. S. Sams, and A. Zahra, "Multimodal music emotion recognition in Indonesian songs based on CNN-LSTM, XLNet transformers," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 1, pp. 355–364, 2023.
[19] Y. Qiao, J. Mu, J. Xie, B. Hu, and G. Liu, "Music emotion recognition based on temporal convolutional attention network using EEG," *Frontiers in Human Neuroscience*, vol. 18, 1324897, 2024.
[20] M. Zhang, Y. Zhu, W. Zhang, Y. Zhu, and T. Feng, "Modularized composite attention network for continuous music emotion recognition," *Multimedia Tools and Applications*, vol. 82, no. 5, pp. 7319–7341, 2023.
[21] F. Eyben, and B. Schuller, "openSMILE: The Munich open-source large-scale multimedia feature extractor," *ACM SIGMultimedia Records*, vol. 6, no. 4, pp. 4–13, 2015.
[22] S. H. Deshmukh, and S. Bhirud, "North Indian classical music's singer identification by timbre recognition using MIR toolbox," *International Journal of Computer Applications*, vol. 91, no. 4, 2014.
[23] B. Bhattarai, and J. Lee, "Automatic music mood detection using transfer learning and multilayer perceptron," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 19, no. 2, pp. 88–96, 2019.
[24] E. Aldhahri, "The Use of Recurrent Nets for the Prediction of e-Commerce Sales," *Engineering, Technology and Applied Science Research*, vol. 13, no. 3, pp. 10931–10935, 2023.
[25] W. Lu, J. Li, J. Wang, and L. Qin, "A CNN-BiLSTM-AM method for stock price prediction," *Neural Computing and Applications*, vol. 33, no. 10, pp. 4741–4753, 2021.