Scholar Publishing Group
International Journal of Art Innovation and Development
https://doi.org/10.38007/IJAID.2022.030405
ISSN 2789-7192 Vol. 3, Issue 4: 53-60

Music Emotion Recognition Model Integrating Deep Learning

Huimin Yang*

Hebei Chemical & Pharmaceutical College, Shijiazhuang, China

**corresponding author*

Keywords: Deep Learning, Music Emotion Recognition, Recognition Model, CNN-SVM Model

Abstract: With the development of digital music technology, people begin to explore new classification and recognition methods to retrieve target music from massive data. Music is the carrier of emotion, and the recognition research based on music emotion classification has very important objective significance. The purpose of this paper is to study a music emotion recognition model incorporating deep learning. The audio features of music are extracted and screened based on the underlying audio features, fused with the audio features obtained by deep learning, and combined with the CNN-SVM network model for music emotion classification and recognition. The advantages of the two are combined to carry out the task of music emotion classification, and the final comparative experiments are carried out on three different datasets. Experiments show that the CNN-SVM model in this paper, combined with the filtering of the CNN layer and the new chord vector feature, achieves the best results on all three datasets.

1. Introduction

Emotion or emotion is a part of human subjective attitude. It is a complex and stable attitude experience and evaluation of human beings after receiving external stimuli, and is also affected by human values and moral values [1-2]. The use of external stimuli to induce emotions or emotions is of great significance to our study of emotion generation. Music is an important carrier for human beings to try to express what human beings think and feel. There are countless great musicians in history who have used music to carry their mental journeys. It can be seen that music has a special stimulating effect on human emotions [3-4].

Deep learning models have been successfully used in emotion recognition. Veltmeijer EA studies the disentanglement for the primary task of the secondary task of facial recognition. A multi-task framework was developed to extract low-dimensional embeddings designed to capture emotion-specific information [5]. Emotion Recognition has been one of the most fascinating topics recently. Demircan S extracts MFCCs (Mel Frequency Cepstral Coefficients) from EmoDB. Obtain various statistical values from MFCC. ANN and 10 cross validations were used for classification. A close understanding of the three emotions is achieved in the app. As a result, it can be seen that the classification accuracy is improved [6]. Learning the emotional information contained in the melody of songs to identify the emotional types of other songs has important practical value for the research of emotional computing [7].

The structure of convolutional neural network is improved, and the support vector machine is combined with convolutional neural network to realize the classification of music emotion. In order to conduct a more comprehensive learning and classification of emotional features, this paper builds a convolutional neural network to learn the input features of the network. While reducing the dimension of the data, the emotional features are learned at a deeper level. Although support vector machines cannot efficiently extract high-latitude features. Therefore, this paper more effectively achieves the classification of music emotion.

2. Research on Music Emotion Recognition Model Integrating Deep Learning

2.1. Music Emotion Classification Model

The goal of developing a music emotion classification system is to map the music emotion feature data into the four main emotions identified by the classification system to understand the automatic recognition of emotion [8-9]. The classification process generally has two stages: the first stage is the training stage, that is, the labeled training set is sent to a very emotional classifier. The second Step is the testing step, i.e. select a new music file as the test set, predict it using the trained classification method, compare the predicted result with the actual set, use the statistical sensitivity value as 'one of the analysis' to evaluate the model performance parameter [10].

The SVM algorithm has great advantages in dealing with the problem of few samples and nonlinearity. Therefore, the SVM classification algorithm is used to construct a multi-classification model of music emotion to identify music emotion. SVM has a solid theoretical foundation, mainly including the characteristics of less computation, strong data generalization performance, and easy implementation [11-12].

2.2. Convolutional Neural Network CNN

CNN is a multi-layer network structure, mainly composed of convolutional layer, pooling layer and fully connected layer. Compared with ordinary fully connected neural network, the nodes of convolutional neural network are partially connected. connected [13-14]. The convolutional layer and the pooling layer are alternately connected to achieve layer-by-layer feature extraction, and finally the fully connected layer completes the classification task. The convolution layer can extract the features of the local area of the previous layer through the convolution operation. In image processing, the convolution calculation of an image is essentially the filtering process of the image through the convolution kernel. The image convolution calculation process can be expressed by formula (1) [15].

$$f(x,y)*w(x,y)=\sum_{s=-a}^{a}\sum_{t=-b}^{b}w(s,t)f(x-s,y-t) \quad (1)$$

Among them, f(x, y) represents the gray value of the point on the xth row and the yth column on the image, and the convolution kernel is equivalent to a weight template. It slides and walks in the image matrix, and once slides, a convolution is performed. Calculate and take the result as the response of the corresponding pixel on the image [16].

2.3. Audio Feature Extraction

Before feature extraction, it is necessary to perform format conversion, pre-emphasis and other processing on audio, and then perform feature extraction on Mel Frequency Cepstral Coefficients (MFCC) [17-18]. The specific process is as follows:

(1) Convert all song formats to wav format;

(2) Pre-emphasize the audio to improve the high-frequency resolution in the music signal, which is achieved by adding a filter, using the formula as shown in formula (2):

$$H(n) = 1 - \mu n^{-1} \quad (2)$$

Among them, n represents the input signal, and formula (2) is transformed into a functional equation, which is formula (3):

$$y(n) = x(n) - \mu * x(n-1) \quad (3)$$

The role of the pre-emphasis factor μ is to multiply with the frequency domain.

(3) Extract MFCC features. The sampling points are converted into spectrum and energy distribution through Fourier transform, the formula is shown in (4):

$$x(m) = \sum_{m=0}^{N-1} x(n)e - j\frac{2\pi nk}{N}, 0 \le m \le N-1 \quad (4)$$

Among them, n represents the frame value, and N is the number of discrete sampling points corresponding to the audio frame used in the Fourier transform. Transform the music signal from the frequency domain to the cepstral frequency domain, as shown in equation (5):

$$C(n) = m = \sum_{m=0}^{N-1} s(m)\cos\left(\frac{\pi n(m-0.5)}{M}\right), n = 1,2,\ldots,L \quad (5)$$

Among them, M is the number of Mel filters, n=1, 2,..., L is the order of MFCC.

3. Investigation and Research on Music Emotion Recognition Models Integrating Deep Learning

3.1. Dataset Introduction

The EMA dataset consists of 1654 pure music pieces in 4 emotion categories: cheerful; excited; nervous; joyful. In the research process, in order to improve the processing convenience, the first 50 seconds of each song was simply selected, and the zero-fill operation was performed if the duration was less than one minute.

The Emotion dataset consists of 1677 pieces of music in MP3 format, and music emotions are classified into 4 categories: anger, happiness, relaxation, and sadness. The length of the music varies from 40 seconds to 60 seconds. For the convenience of the experiment, the WAV format conversion is performed, and only the first 40 seconds of each music is used, and the zero-filling operation is performed if it is less than 40 seconds.

The 4Q-emotion dataset consists of 1471 pieces of music in MP3 format. Music emotions are not classified according to emotional words, but according to the four labels of Q1, Q2, Q3 and Q4. For WAV format conversion, only the first 40 seconds of each music is used, and the zero-fill operation is performed if it is less than 40 seconds.

3.2. Construction of Classification Model Based on CNN-SVM

This paper proposes a fusion classification algorithm, which uses the convolutional neural network to extract features, and then uses the SVM classifier to classify, that is to say, the output of the last fully connected layer of CNN is used as the input of the SVM classifier to establish a new classification model. First, the processed music and text information is used as the input of the CNN network model. The CNN learns two modalities of input data to obtain the musical features and textual features extracted by the CNN. In the fully connected layer of the network, the features of different modalities are fused, and the fused multimodal features are used as the input of the classifier, that is to say, the SVM classifier is used to test and classify the fused multimodal features. While retaining the deep features of different modalities, a better music emotion classification effect can be obtained.

It can be seen that the fully connected layer of the convolutional neural network is combined with the support vector machine classifier, and the output of the network is used as the input of the classifier. CNN-SVM adopts the stochastic gradient descent algorithm, and the convolutional neural network is composed of multi-layer sensors. The model description of the output of the multi-layer sensors is shown in formula (6):

$$f(x) = (w * \phi(x) + b) \quad (6)$$

W is the weight, b is the offset vector, and the variable ϕ represents other parameters. The support vector machine used in CNN-SVM is a linear support vector machine.

4. Analysis and Research of Music Emotion Recognition Model Integrating Deep Learning

Table 1, Table 2 and Table 3 show the results obtained by each model in the EMA data set, Emotion data set and 4Q. The specific analysis of the content of this table shows that the accuracy of music emotion classification is The EMA dataset reaches 85%, which is 17% higher than CNN, 35% higher than SVM, and 8% higher than LSTM, as shown in Figure 1; the classification accuracy of the model in the Emotion dataset reaches 88%, which is higher than CNN improves by 22%, 31% better than SVM, and 17% better than LSTM, as shown in Figure 2.

Table 1. Model classification comparison (EMA)

Model	Accuracy	Precision	Recall	F1	training time (s)
CNN	0.68	0.70	0.71	0.71	201.05
SVM	0.50	0.59	0.57	0.58	185.04
LSTM	0.77	0.72	0.78	0.81	316.67
CNN-SVM	0.85	0.89	0.80	0.81	305.14

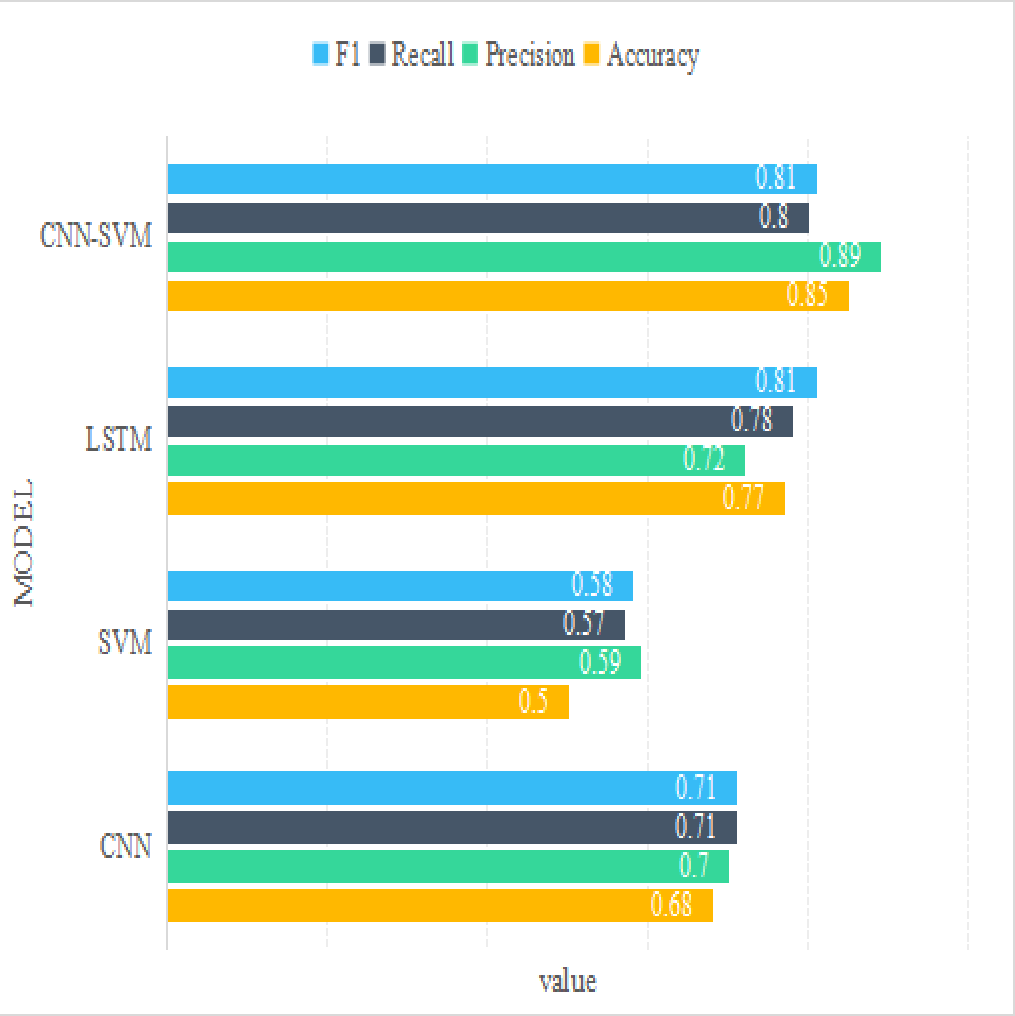

Figure 1. Model comparison results in the EMA dataset

Table 2. Model classification comparison (Emotion)

Model	Accuracy	Precision	Recall	F1	training time (s)
CNN	0.66	0.65	0.68	0.69	204.06
SVM	0.57	0.55	0.59	0.60	100.14
LSTM	0.71	0.77	0.76	0.70	245.57
CNN-SVM	0.88	0.81	0.85	0.84	200.14

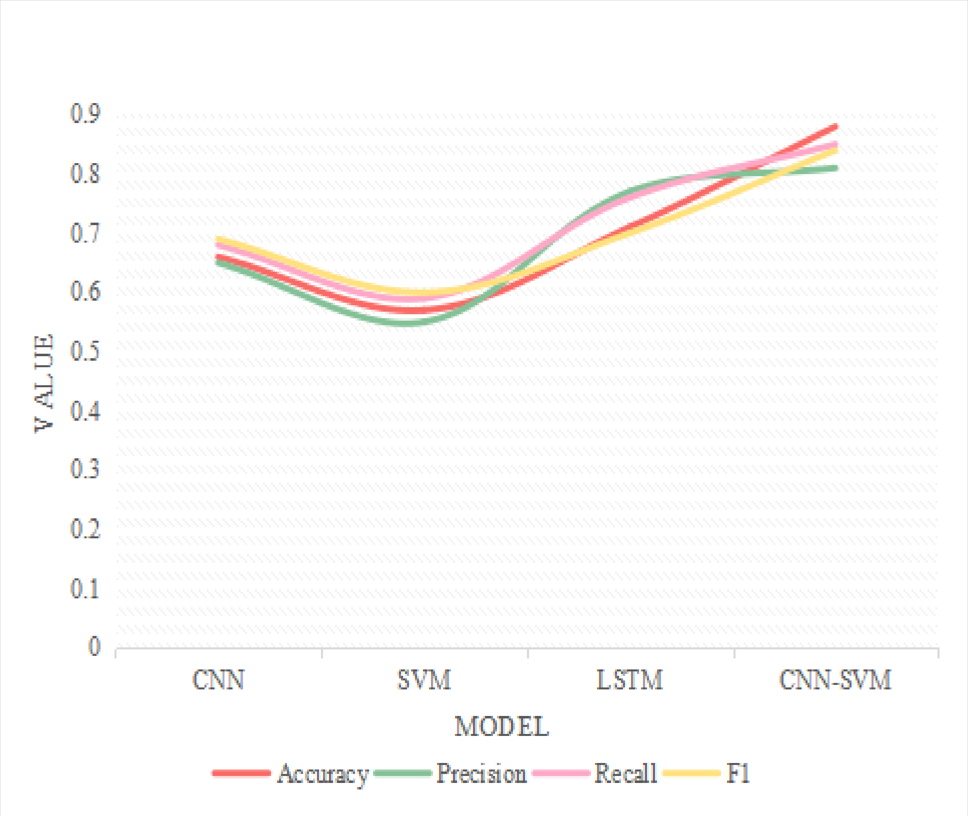

Figure 2. Model comparison results in the Emotion dataset

Table 3. Model classification comparison (4Q)

Model	Accuracy	Precision	Recall	F1	training time (s)
CNN	0.74	0.78	0.80	0.76	162.19
SVM	0.70	0.71	0.75	0.75	87.42
LSTM	0.85	0.88	0.84	0.84	187.07
CNN-SVM	0.90	0.92	0.90	0.91	100.94

It can be seen that on different datasets, the CNN-SVM-based model is much more efficient than the CNN-based model training. This is because the model depth of CNN-SVM is much shallower than that of CNN, which can greatly reduce the complexity of the model. At the same time, the difference between the accuracy of the SVM model and the CNN model is only about 10%. The random number of CNN makes up for the shortcomings of SVM's inability to extract depth information to a certain extent, so the CNN model is better than the SVM model and training efficiency. When combining CNN with SVM, although the model training efficiency is not high, the classification accuracy is greatly improved.

5.Conclusion

This paper first proposes some improvements to image analysis techniques by studying the fundamentals of digital music and combining them with relevant data. Then, a music emotion recognition model is established, adjusted according to the amount of data, and verified by corresponding experiments. However, due to the limitation of time and conditions, this paper only analyzes and models the music files in a single format. In the follow-up research, the performance of the model needs to be improved from the following aspects. Although the support vector machine algorithm has strong adaptability, it also has problems such as error accumulation and poor effect on unbalanced data. In the future, it is necessary to study other combined models of support vector machines to make the model have better fault tolerance; or to establish another classification model to make the recognition model more accurate and efficient. In the future, the recognition model can be combined with more fields to explore more possibilities.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Alghifari M F, Gunawan T S, Kartiwi M. Speech emotion recognition using deep feedforward neural network. Indonesian Journal of Electrical Engineering and Computer Science, 2018, 10(2):554-561.

[2] Jain N, Kumar S, Kumar A, et al. Hybrid deep neural networks for face emotion recognition. Pattern Recognition Letters, 2018, 115(NOV.1):101-106.

[3] Kshirsagar P. Face And Emotion Recognition Under Complex Illumination Conditions Using Deep Learning With Morphological Processing. Journal of Interdisciplinary Cycle Research, 2021, XIII(VI):324-331.

[4] Samadiani N, Huang G, Hu Y, et al. Happy Emotion Recognition From Unconstrained Videos Using 3D Hybrid Deep Features. IEEE Access, 2021, PP(99):1-1.

[5] Veltmeijer E A, Gerritsen C, Hindriks K. Automatic emotion recognition for groups: a review. IEEE Transactions on Affective Computing, 2021, PP(99):1-1.

[6] Demircan S, Kahramanli H. Application of ABM to Spectral Features for Emotion Recognition. Mehran University Research Journal of Engineering and Technology, 2018, 37(4):452-462.

[7] Mesnyankina K K, Anishchenko S I, Kalinin K B. The Correlation Between the Set of Mental Functions and Emotion Recognition Skills Formation in Children with Autism Spectrum Disorder. Autism and Developmental Disorders, 2020, 18(4):13-22.

[8] Shukla S, Jain M. A novel stochastic deep conviction network for emotion recognition in speech signal. Journal of Intelligent and Fuzzy Systems, 2020, 38(2):1-16.

[9] Schmidt T, Schlindwein M, Lichtner K, et al. Investigating the Relationship Between Emotion Recognition Software and Usability Metrics. i-com, 2020, 19(2):139-151.

[10] Lotfalinezhad H, Maleki A. Application of multiscale fuzzy entropy features for multilevel

subject-dependentemotion recognition. Turkish Journal of Electrical Engineering and Computer Sciences, 2019, 27(6):4070-4081.

[11] Ozseven T. A novel feature selection method for speech emotion recognition. Applied Acoustics, 2019, 146(MAR.):320-326.

[12] Kwak Y J, Lee H S. A Study on Emotion Recognition of Children with ADHD through Computerized Facial Morphing Task. JOURNAL OF SPECIAL EDUCATION & REHABILITATION SCIENCE, 2018, 57(4):41-56.

[13] Aishwarya R. Feature Extraction for Emotion Recognition in Speech with Machine Learning Algorithm. International Journal of Advanced Trends in Computer Science and Engineering, 2020, 9(4):4998-5002.

[14] Tehmina K, Muhammad A S, Muhammad M, et al. Emotion recognition from facial expressions using hybrid feature descriptors. IET Image Processing, 2018, 12(6):1004-1012.

[15] Kaya H, Karpov A A. Efficient and effective strategies for cross-corpus acoustic emotion recognition. Neurocomputing, 2018, 275(JAN.31):1028-1034.

[16] Albraikan A, Tobon D P, Saddik A E. Toward User-Independent Emotion Recognition Using Physiological Signals. IEEE Sensors Journal, 2018, PP(99):1-1.

[17] Nakisa B, Rastgoo M N, Tjondronegoro D, et al. Evolutionary computation algorithms for feature selection of EEG-based emotion recognition using mobile sensors. Expert Systems with Applications, 2018, 93(mar.):143-155.

[18] Ho N H, Yang H J, Kim S H, et al. Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention based Recurrent Neural Network. IEEE Access, 2020, PP(99):1-1.