Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 4606027, 11 pages
https://doi.org/10.1155/2020/4606027

Research Article

A Multimodal Music Emotion Classification Method Based on Multifeature Combined Network Classifier

Changfeng Chen and Qiang Li

Institute of Intelligent and Software Technology, Hangzhou Danzi University, Hangzhou 310018, China

Correspondence should be addressed to Qiang Li; hzlee@hdu.edu.cn

Received 25 May 2020; Accepted 29 June 2020; Published 1 August 2020

Academic Editor: Piotr Jędrzejowicz

Aiming at the shortcomings of single network classification model, this paper applies CNN-LSTM (convolutional neural networks-long short-term memory) combined network in the field of music emotion classification and proposes a multifeature combined network classifier based on CNN-LSTM which combines 2D (two-dimensional) feature input through CNN-LSTM and 1D (single-dimensional) feature input through DNN (deep neural networks) to make up for the deficiencies of original single feature models. The model uses multiple convolution kernels in CNN for 2D feature extraction, BiLSTM (bidirectional LSTM) for serialization processing and is used, respectively, for audio and lyrics single-modal emotion classification output. In the audio feature extraction, music audio is finely divided and the human voice is separated to obtain pure background sound clips; the spectrogram and LLDs (Low Level Descriptors) are extracted therefrom. In the lyrics feature extraction, the chi-squared test vector and word embedding extracted by Word2vec are, respectively, used as the feature representation of the lyrics. Combining the two types of heterogeneous features selected by audio and lyrics through the classification model can improve the classification performance. In order to fuse the emotional information of the two modals of music audio and lyrics, this paper proposes a multimodal ensemble learning method based on stacking, which is different from existing feature-level and decision-level fusion methods, the method avoids information loss caused by direct dimensionality reduction, and the original features are converted into label results for fusion, effectively solving the problem of feature heterogeneity. Experiments on million song dataset show that the audio classification accuracy of the multifeature combined network classifier in this paper reaches 68%, and the lyrics classification accuracy reaches 74%. The average classification accuracy of the multimodal reaches 78%, which is significantly improved compared with the single-modal.

1. Introduction

Music contains a wealth of human emotional information, and the study of music emotion classification is helpful for the organization and retrieval of massive music data. With the rise of artificial intelligence technology, computers can realize functions of complex emotion analysis and calculation, and scholars' research on music emotion feature extraction and classification models are also gradually launched [1]. Multiple emotional features with heterogeneity can be extracted from music, audio features mainly include spectrograms and LLDs, and lyrics features primarily contain word embedding and word frequency vector. However, it is difficult to combine two categories of heterogeneous features of audio or lyrics through a single network classifier, and the classification effect is constrained. And, related researches mainly focused on single-modal analysis of audio or lyrics, ignoring the correlation between the two modals and certain emotion information are missing to some degree. This paper constructs a multimodal music emotion classification system through a multifeature combined network classifier, which can effectively enhance the classification performance.

In the music emotion classification research of audio, Hwang et al. [2] extracted 37 features to represent music samples, including rhythm, dynamics, and pitch, and utilized K-nearest neighbor classifier to output the results. Zhang et al. [3] extracted 8 kinds of acoustic features to

represent the arousal dimension in the 2D music emotion model and applied logistic regression methods to explain these features, but this research mainly focused on the 1D emotion and did not verify the specific emotion category. Ramani and Priya [4] extracted Mel frequency, spacing, and zero-crossing rate as separate representations of the best-fit ratio and used genetic algorithm as the classification technique. This research verified the good classification ability of frequency-domain features on emotion information. Chen et al. [5] proposed a music emotion detection system based on deep Gaussian process, which classified 9 emotions through 9 classifiers. Lin et al. [6] proposed a two-layer SVM (support vector machine) sentiment classification structure based on the relationship between music genre and emotion to utilize the genre information available in music tags. These machine learning methods mostly used acoustic or frequency features for classification, neglecting the features of the spectrogram and sequence. Seo and Huh [7] used different classification algorithms such as random forest, DNN, and K-nearest neighbor for comparative analysis and used SVM as the best classification method. Yaxin [8] used BP neural network to classify emotion features. Zhao and et al. [9] proposed a music emotion classification method based on RNN (recurrent neural networks). The above researches validated the representation of different audio features for emotional information and the applicability of single classification method in music emotion classification.

In the music emotion classification research of lyrics, An [10] applied the emotion dictionary to construct the features of lyrics and adopted a simple Bayesian classifier for classification verification; the accuracy reached 68%. Chen and Tang [11] used the TF-IDF word frequency statistical method to extract the features of the lyrics text and constructed a composite emotion point matrix for each song for further classification. He et al. [12] constructed the lyrics text feature representation through the n-gram language model, using supervised learning methods, naive Bayes, and support vector machines to check the classification performance. Lee et al. [13] used Word2vec to extract the word embedding of the lyrics text and used DNN to train the data under supervised learning. Reddy and Mamidi [14] constructed a lyrics classification model with mixed features and adopted LSTM for classification, with an accuracy of 76.6%. Wang and Zhao [15] proposed a CNN-based pretrained word embedding model which can effectively extract the emotional features of Chinese lyrics compared with traditional classification models. The above researches extracted the emotional information of lyrics through different text feature representation methods and selected the single classifier to output the results.

The CNN-LSTM combined network classification method has launched a series of applications in the fields of the audio and text classification. Satt et al. [16] studied the speech emotion classification effect through CNN-LSTM combined model using the spectrogram less than 3 s. Kim and Saurous [17] adopted 20 features in the eGeMAPS feature set and CNN-LSTM combined model for speech emotion recognition. Gang and Liu [18] proposed an AC-BiLSTM text classification architecture, which included BiLSTM, attention model, and convolutional layer. Experiments showed that this combined network architecture performed significantly better than other classification methods. Wang et al. [19] also used convolutional recurrent neural networks for text classification verification and concluded that this method has a good accuracy in text classification. Zheng and Zheng [20] proposed the BRCAN model, which combined BiLSTM and CNN with attention model and Word2vec to achieve fine-grained text classification. The above researches showed that, compared with single network classification model, CNN-LSTM combined network can use CNN's ability to extract 2D feature and LSTM's ability to process sequence data. CNN-LSTM combined network has better effects in audio and text classification and can be applied in the field of music emotion classification.

Multimodal music fusion helps to boost emotion classification performance. Mihalcez and Strapparava [21] extracted lyrics features through the bag-of-words model, audio features through rhythm, timbre, etc. and combined them into multimodal feature representations. After linear regression comparison and verification, the fusion features well improved the classification performance. Panda et al. [22] extracted melody features and semantic features from MIDI music files and lyrics, respectively, and used supervised learning methods for classification verification. The experimental results showed that the best performance of classification was 44.3% when only using standard audio functions, and with the use of multimodal features, the rate increased to 61.1%. Su et al. [23] adopted the AdaBoost method with decision tree to achieve multimodal music emotion classification. Rachman et al. [24] used a random forest method to fuse multimodal information of music audio and lyrics. Shi and Feng [25] proposed a decision-level fusion method of music based on improved LFSM algorithm. Su and Xue [26] extracted descriptive sentence-level lyrics and audio features from music, adopted a graph-based multimodal classification model for music emotions, and aggregated the obtained emotion category predictions for each music sentence through a simple voting scheme. Generally, existing multimodal music emotion fusion methods can be divided into decision fusion and feature fusion, both of which have certain shortcomings, and the effects of emotion classification are limited.

Based on the above research status, this paper proposes a multimodal music emotion classification method, and three main contributions are shown as follows:

Apply CNN-LSTM combined network in the field of music emotion classification, and propose a multi-feature combined network classifier. Aiming at the limitation of a single feature, the model improves existing CNN-LSTM classification models and uses 2D + CNN-LSTM and 1D + DNN architecture to combine two types of heterogeneous features to make up for the lack of classification performance.

Propose a multimodal ensemble learning method based on Stacking. Aiming at the shortcomings of traditional feature-level and decision-level multimodal fusion, the

method adopts different modal classification models as the basic classifier, outputs emotional label results to form a new dataset, and uses sub classifier to fuse emotion output so as to effectively solve the problem of heterogeneity as well as significantly improve the classification accuracy compared with that of the single-modal.

Propose a preprocessing method of music audio for optimizing the dataset. According to the differences in the duration and composition between music audio and speech, the preprocessing method finely segments the audio clips and extracts pure background sound through human voice separation to optimize the dataset and improve the classification performance.

2. Multifeature Combined Network Classifier

2.1. Previous CNN-LSTM Classification Model. CNN-LSTM combined network has been used for a series of single-modal applications in speech and text classification. Satt [16] studied the speech emotion classification through the combined model using the spectrogram of less than 3 s as input; the model is as shown in Figure 1(a). Wang et al. [19] also used CNN-LSTM model for long text classification, as shown in Figure 1(b).

The above classification models verify the classification abilities of CNN-LSTM combined network in audio and text, and can be applied in the field of music emotion classification. However, there are still some issues of applicability: Compared with speech, real music audio is longer and more complex in terms of composition, requiring detailed preprocessing. CNN-LSTM can also add the attention model to further optimize performance. At the same time, the above two models only use 2D emotional features such as spectrogram and Word2vec, which have certain limitations. Many 1D features with certain emotion classification performance, such as audio LLDs and text word frequency vector, have been used in a large number of studies. Combining 2D and 1D features can effectively make up for the deficiencies of the single feature.

2.2. Multifeature Combined Network Classifier Based on CNN-LSTM. According to the theme of music emotion classification, the audio features extracted from music contain two types: spectrogram (2D) and LLDs (1D), and the features of lyrics text include two types: word embedding (2D) and word frequency vector (1D). Among them, the spectrogram needs to simultaneously integrate the spectral characteristics and the sequence characteristics, and the word embedding has high dimensions and sparseness. These 2D features cannot be well classified when only using a single network classification model. Due to the heterogeneity of the two types of features, if the 2D features and 1D features are directly combined at the input layer, dimension reduction and normalization are required, which brings part of emotional loss and reduces the classification accuracy. Therefore, we propose a multifeature combined network classifier by improving the CNN-LSTM classification model, mainly composed of two parts: 2D + CNN-LSTM and 1D + DNN. Through this model, the two types of features can concatenate and finally output the single-modal emotion classification results, as shown in Figure 2.

The model can be used as a general classification model for multifeature classification output. In this paper, by inputting the emotional feature of different music modal, we, respectively, construct single-modal classifiers of audio and lyrics. Specifically, for the audio classifier, CNN plays a dominant role to extract 2D feature data and requires a deeper convolution operation. The size of the convolution window and stride will influence the classification performance; for the lyrics classifier, the original serialized text feature vectors are difficult to train due to their high dimensionality and sparsity, CNN mainly provides feature compression capabilities, and BiLSTM and attention model have a greater impact on classification accuracy compared with convolutional layers.

2.3. Specific Description of Audio and Lyrics Classifier

2.3.1. Audio Classification Input Layer. Audio features comprise two types: spectrograms and LLDs. The spectrogram is generated by short-time Fourier transform and combined into 2D feature input through timing sequence. LLDs are generally calculated by single-frame audio, HSFs (high-level statistics functions) are features obtained by doing some statistics on the basis of LLDs. According to the theme of music emotion classification and the research concerning audio feature in the literature, the 1D audio features extracted in this paper are shown in Table 1.

MFCC is a cepstrum parameter extracted in the frequency domain of Mel scale and is widely used in automatic speech and speaker recognition [27]. ZCR is the rate of the sign changes along a frame. The spectral centroid and the spectral spread are measures for characterizing the distribution of the frequency components of a signal. The spectral rolloff is a measure of the skewness of the spectrum. The spectral flu indicates how quickly the spectral information of a signal is changing. Chroma feature is the collective name of chroma vector and chromagram, which has good applications in the field of music chord detection.

Each frame of the spectrogram is normalized to the size of $128 * 128 * 3$ and then input into the CNN layer after the timing combination for the next convolution pooling operation. Extracted to the maximum, mean, and variance combination as HSF features with a dimension of 20, LLDs are used as another feature representation of audio samples and is input into DNN for the next operation. Audio classification input is shown in Figure 3.

2.3.2. Lyric Classification Input Layer. Lyric input is divided into two features: word embedding and word frequency vector. Among them, the word embedding is extracted from Word2vec with a dimension of 100, which is input into the CNN layer for feature compression extraction.

For word frequency vector, the chi-squared test feature extraction method is derived from the CHI test method in

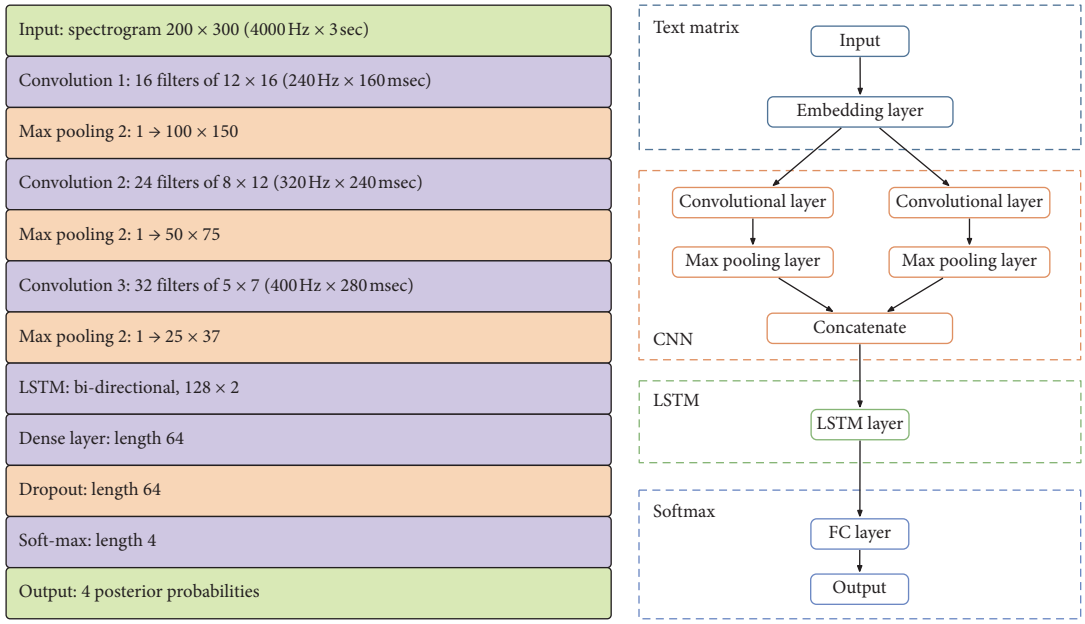

Figure 1: Previous CNN-LSTM classification model. (a) Speech classification model proposed by Satt. (b) Text classification model proposed by Wang.

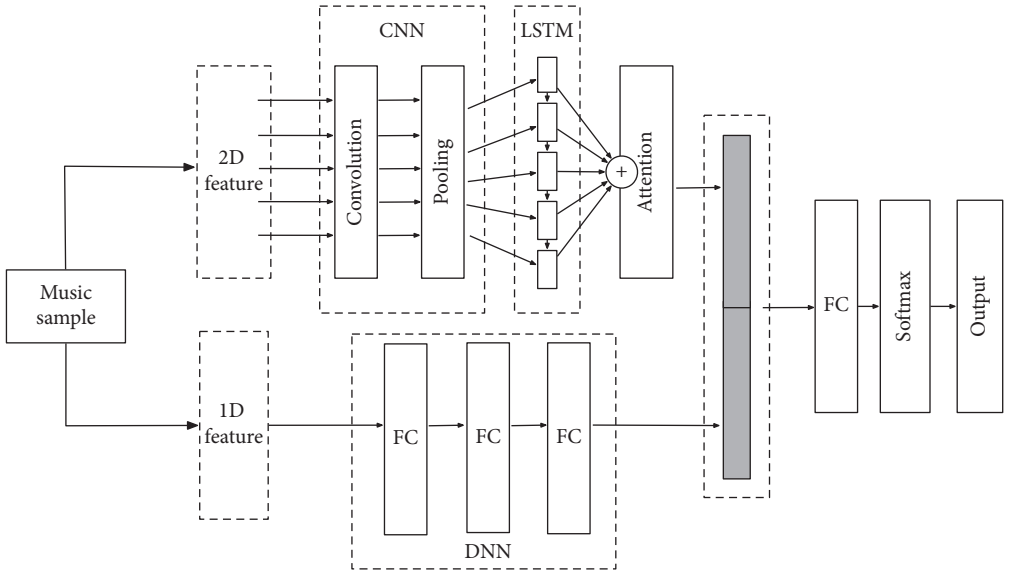

Figure 2: Multifeature combined network classifier based on CNN-LSTM.

Table 1: 1D audio features.

Category	Features
LLDs	Mel frequency cepstrum coefficient (MFCC), zero crossing rate (ZCR), spectral centroid, spectral spread, spectral rolloff, spectral flu, and chroma features
HSFs	Maximum, mean, and variance

mathematical statistics, which is better than the TF-IDF method though SVM in emotion classification [28] and is used to characterize the correlation between two random variables. In the performance of lyrics text, a large number of compact description words often appear for the lyrics of a specific emotion type. Statistical processing of these special

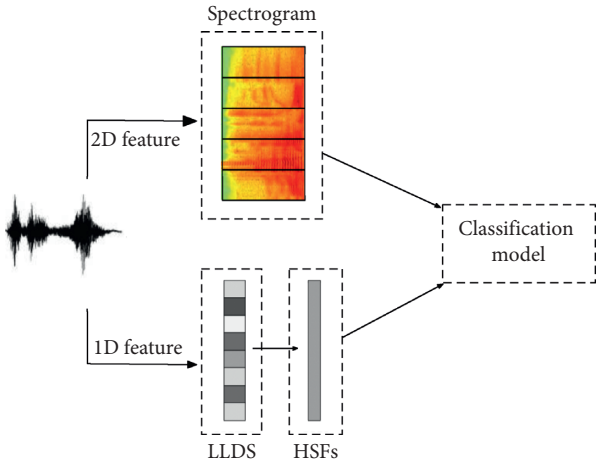

Figure 3: Audio classification input.

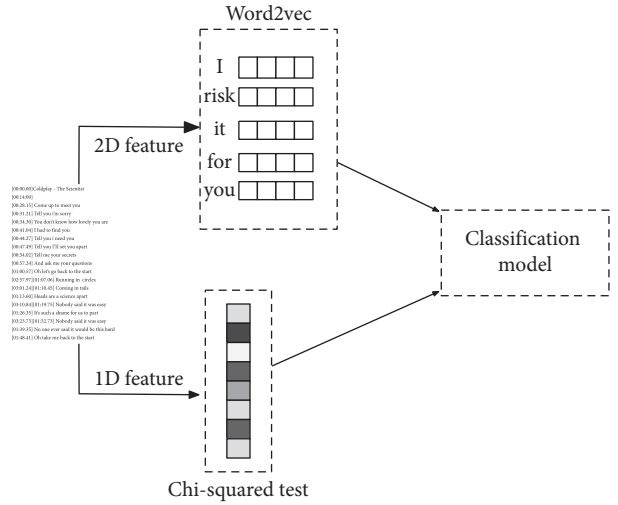

Figure 4: Lyrics classification input.

emotion words can improve the performance of lyrics text classification. This paper uses the chi-squared test as the 1D feature representation of the lyrics and is input into the DNN layer for the next operation. The lyrics classification input is shown in Figure 4.

2.3.3. CNN Layer. In order to deal with 2D features, a multiscale convolution kernel is used in the CNN layer to perform a convolution operation on the input data, then a pooling operation is used to further extract features, and finally the network output results are merged into a timing serialized representation.

The CNN layer contains 2 convolutional layers and 2 pooling layers. The first layer of convolution input is an audio spectrogram or lyrics Word2vec, and the convolution operation is performed by 64 convolution kernels of 2×2, with a step size of 1. Then, the pooling operation uses the max pooling layer to pool the convolution results. The second layer of convolution process is the same as that of the first layer, and the convolution kernel size is adjusted to 3×3.

During the convolution process of the first layer, a single convolution kernel is firstly used to calculate each local feature of the input, as shown in equation (1), where W_F represents the convolution kernel with height F, X represents the input vector, and b is a constant. Then, the calculated features are connected vertically, as shown in equation (2), where H represents the height of the feature map after convolution. Finally, the calculation result is nonlinearly calculated by the *relu* activation function to obtain the final convolution feature, as shown in Equation (3):

$$h_{1F}(i) = f\left(W_F \cdot X(i\colon i + F - 1) + b\right), \tag{1}$$

$$h_{1F} = \left[h_{1F}(1), h_{1F}(2) \cdots h_{1F}(H)\right], \tag{2}$$

$$h_{r1F} = \text{relu}\left(h_{1F}\right). \tag{3}$$

During the pooling process, the max pooling operation is taken, as shown in equation (4), and the size of the window changes with the length of the sample. In order to maintain timing serialized representation, the merged result is connected to the input of the LSTM layer, as shown in equation (5):

$$h_{rP1F} = \max\left(h_{r1F}\right), \tag{4}$$

$$h_1 = \text{Concatenate}\left(h_{rP1F_1}, h_{rP1F_2} \cdots\right). \tag{5}$$

2.3.4. BiLSTM and Attention Layer. LSTM can effectively capture the context information of the input sequence and solve the problem of preservation and transmission of serialized information. The feature sequence output from the CNN layer can extract the features of each moment through the LSTM unit.

The BiLSTM layer in the model has 128 units, and the output can be expressed as $[r_{(1)}, r_{(2)}, r_{(3)}, \ldots, r_{(N)}]$, where N is the number of units in the layer.

When LSTM performs many-to-one classification tasks, the output at the last moment is expressed as the classification output result. After introducing the attention model, important information can be captured, and all LSTM output can be weighted and summed according to the weight as the final output representation. For each vector r_i output by the BiLSTM, its attention weight value a_i can be calculated by the following formula:

$$a_i = \frac{\exp\left(f\left(r_i\right)\right)}{\sum_j \exp\left(f\left(r_j\right)\right)}, \tag{6}$$

where $f(r_i)$ is the score function (Softmax).

The output result of the attention model layer att_n is the weighted sum of the attention values of the entire sequence, as shown in the following equation:

$$att_n = \sum_i a_i r_i. \tag{7}$$

2.3.5. DNN Layer. DNN can be understood as a neural network with multiple FC (fully connected layers), which can process 1D features well. The DNN layer in the model contains 3 hidden layers with 256, 128, and 64 nodes, respectively, which are used to synthesize feature information. The input audio HSF feature or lyrics chi-squared test vector is further compressed through the DNN layer.

2.3.6. Output Layer. The output layer is composed of FC and Softmax. The 2D feature output through the CNN layer, LSTM layer, and attention model layer, and the 1D feature output through the DNN layer is concatenated as the final classification feature representation, and then the single-modal emotion classification results of music audio and lyrics are obtained through respective output layers.

3. Multimodal Fusion

The fusion of emotional information of music audio and lyrics can efficiently enhance the classification performance. Multimodal fusion methods in existing research generally contains two types: feature fusion and decision fusion. For feature fusion methods, different modal feature vectors have large heterogeneous differences, if a concatenate method is used to fuse features directly, part of the emotional information will be lost in the dimensionality reduction process, resulting in classification performance degradation. For decision fusion methods, the most common method is linear probability fusion method which only considers the output emotional probabilities of different modals, ignoring the correlation between the features of modals and results in certain limitations. This paper studied the significant problems of feature heterogeneity and feature correlation in multimodal fusion and proposed a stacking ensemble learning method for music emotion classification.

3.1. Stacking Model Building. Stacking is an ensemble learning technique that applies the output of multiple models to yield new models. The core idea is to train the original sample features with different basic classifiers, combine the basic label results obtained by the training as a new dataset sample feature representation, then input it to a subclassifier for learning and training, and finally, output the ensemble classification results. The stacking method has achieved excellent performance in image and text classification tasks. In Chand's research on network attack detection [29], SVM and random forest stacking can provide the best performance, with an accuracy of about 97.50%, which is significantly better than SVM (91.81%). The stacking method is of high efficiency for ensemble of diverse models and can be applied to music emotion classification systems.

In this paper, the audio emotion classifier and the lyrics classifier in Section 2 are used as the basic classifiers to build a multimodal ensemble classification model based on stacking, as shown in Figure 5.

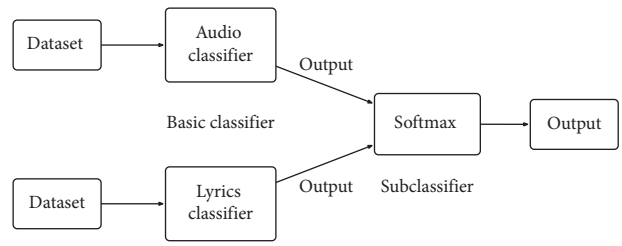

Figure 5: Stacking model for music classification.

3.2. Stacking Model Training. During the training process, the 5-fold cross-validation method was used to solve the problem of overfitting in the stacking process, as shown in Figure 6.

3.2.1. Dataset Processing. The dataset used in this paper, with 2000 samples, is divided into training set and test set with the ratio of 8 : 2. And, on the basis of the original training set, a 5-fold cross-validation is adopted for further division.

3.2.2. Basic Classifier Training. The model contains 2 basic classifiers: multifeature combined network classifier for audio (M1) and multifeature combined network classifier for lyrics (M2). First, the audio classifier (M1) is trained by using the 5-fold cross-validation method on the 4 sets (1280) split from the training set, as shown in Figure 6, and the remaining 1 set (320) is used to predict, as shown in equation (8), where $tr\_1 \sim tr\_5$ represent the original training set and pr_1 represents the new training set:

$$\begin{pmatrix} tr\_1 \\ tr\_2 \\ tr\_3 \\ tr\_4 \end{pmatrix} \longrightarrow^{\text{training}} (tr\_5) \longrightarrow^{\text{test}} (pr\_1). \tag{8}$$

At the same time, trained M1 is predicted on the original entire test set, as shown in equation (9), where te_1 represents the new test set:

$$\begin{pmatrix} tr\_1 \\ tr\_2 \\ tr\_3 \\ tr\_4 \end{pmatrix} \longrightarrow^{\text{training}} (test) \longrightarrow^{\text{test}} (te\_1). \tag{9}$$

Perform the above operations 5 times, respectively, to obtain 5 new pr and te. Connect 5 pr; the size of the final $P1$ is still 1600. And, average the sum of 5 te and the size of the composition $T1$ is 400, as shown in the following equations:

$$(pr\_1) + (pr\_2) + \cdots (pr\_5) \longrightarrow (P1), \tag{10}$$

$$\frac{(te_1) + (te_2) + \cdots (te\_5)}{5} \longrightarrow (T1). \tag{11}$$

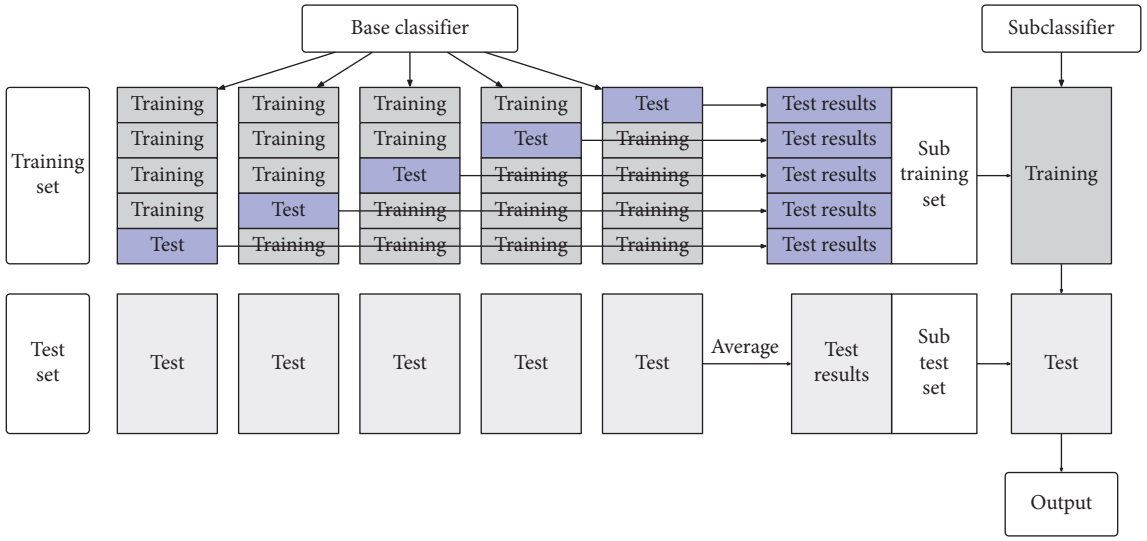

Figure 6: Stacking model training.

For the lyrics classifier (M2), the above operations are also performed to obtain $P2$ and $T2$ and connect $P1$ and $T1$ to form the subtraining set P and the subtest set T (the size is the same as that of the original data set, and the feature includes 2 columns, both are classification labels), as shown in the following equations:

$$(P1) + (P2) = (P1\,P2) \longrightarrow (P), \tag{12}$$

$$(T1) + (T2) = (T1\,T2) \longrightarrow (T). \tag{13}$$

3.2.3. Subclassifier Training. After being trained by the basic classifiers, a new training set P and test set T are generated. The features of different modals are merged into labeling results, eliminating the heterogeneity of the original features. The model uses the Softmax layer as the subclassifier and outputs the final multimodal fusion result, as shown in the following equation:

$$(P) \longrightarrow^{\text{training}} (T) \longrightarrow^{\text{test}} (\text{Output}). \tag{14}$$

In summary, the stacking method ensembles classification model of different modal and integrates the ability of different classifiers to extract features from different angles. In this paper, the stacking-based multimodal ensemble learning method is used to solve the problem of heterogeneity of different modal features. Compared with the feature fusion method, the ensemble results are more stable and accurate and the model program is simple to implement. There is no need to adjust the previously constructed single-modal classification model.

4. Experiments

4.1. Dataset. The dataset used for the experiments in this paper is from the Last.fm tag subset of the million song dataset. According to Thayer's emotion model, four emotional tag music lists are extracted. The emotional tags are angry, happy, relaxed, and sad, and 500 songs are extracted from each emotional list, for a total of 2000. We used script tools to download the song audio and lyrics files in accordance with the tag lists and selected them manually.

4.2. Audio Preprocessing. In order to solve the problem of overlarge feature size and complex composition of real music audio, this paper proposes a preprocessing method, including fine-grained segmentation and vocal separation. The audio samples are preprocessed at four levels to construct 4 experimental datasets, as shown in Figure 7; vote on the output of clips to get emotion classification results.

It can be observed in Figure 7 that the audio effective information after fine-grained segmentation is amplified, and the feature extracted from it is easier to train. Pure background audio clips after vocal separation is more concentrated than the original audio feature. Due to the pause in the singing, the waveform of pure human voice clips has a severe fracture phenomenon.

The classification performance of the four datasets was verified by LLDs features and SVM classifier. The results are shown in Figure 8. The 15 s fine-grained clips have higher accuracies compared with the 30 s clips. The performance of pure human voice clips is poorer, but the pure background audio clips have obtained the highest average classification accuracy, which is consistent with the waveform analysis results. This result is instructive for optimizing the audio dataset to improve the music classification performance. Therefore, the 15 s pure background clips will be selected as the dataset for the audio experiment below.

4.3. Audio Experiment. This group of experiments adopts different classification models to verify audio classification performance. The classification results are shown in Table 2.

The experiment shows that the spectrogram has achieved a certain classification effect through CNN. As its feature

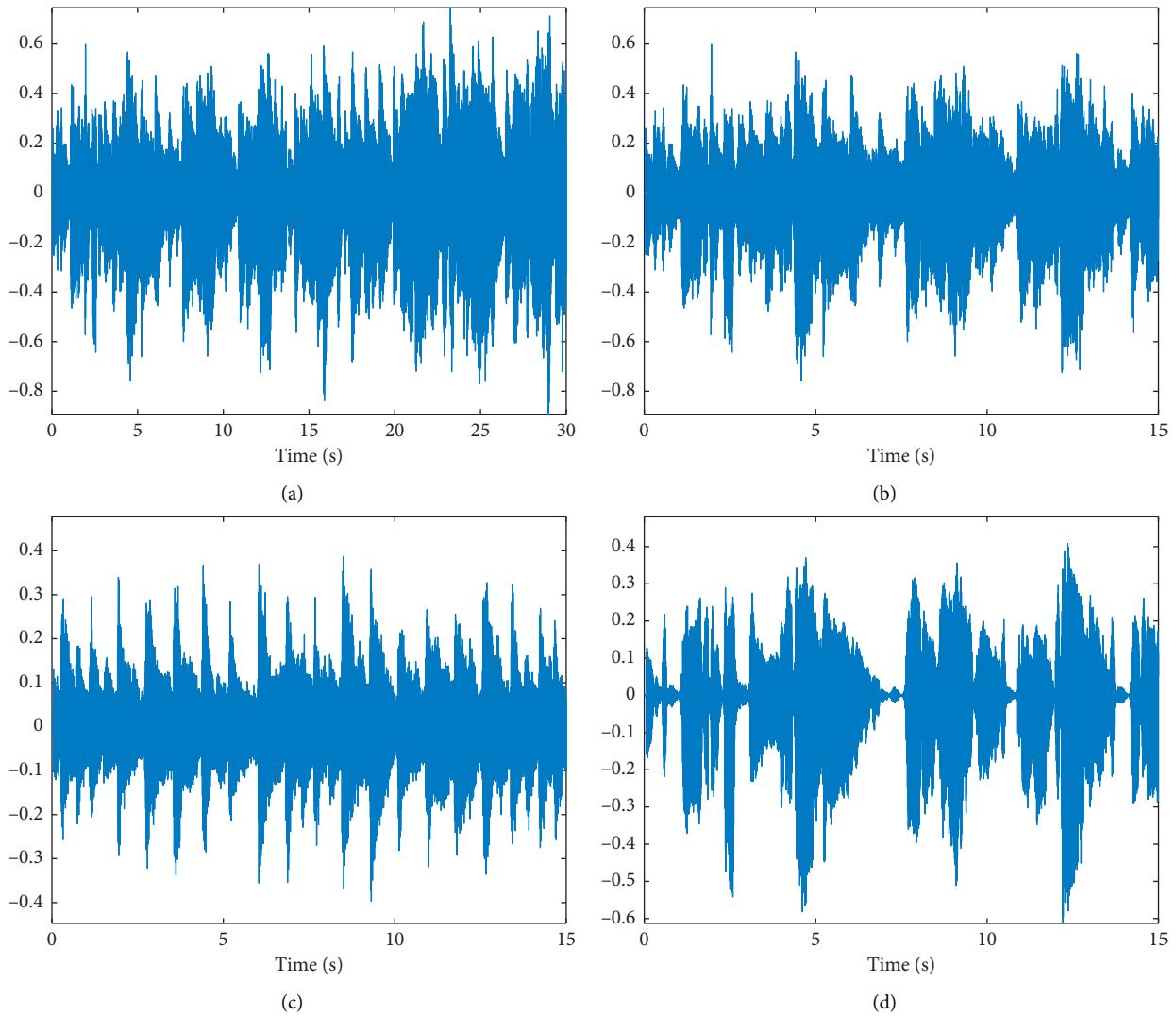

Figure 7: Audio waveform after preprocessing (music from Coldplay-the Scientist). (a) 30 s original; (b) 15 s original; (c) 15 s pure background; (d) 15 s pure human voice.

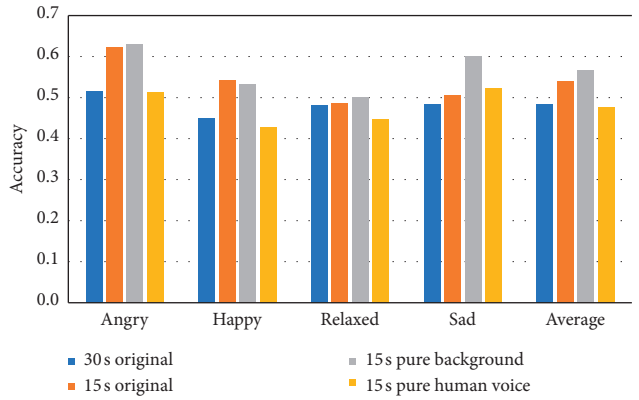

Figure 8: The accuracy of audio classification in 4 preprocessing methods.

dimension is too high, it is not effective to directly adopt the LSTM method for classification. The CNN-LSTM combined network model has good performance compared with that of a single network classification method.

Table 2: Accuracy of different audio classification models.

Classification models	Angry	Happy	Relaxed	Sad	Average
Spectrogram + CNN	0.643	0.594	0.51	0.62	0.592
Spectrogram + LSTM	0.631	0.427	0.54	0.438	0.509
Spectrogram + CNN-LSTM	0.632	0.632	0.632	0.632	0.632
Our model	0.705	0.602	0.688	0.73	0.681

The multifeature combined network classifier proposed in this paper has achieved the best classification effect. The model integrates the spectrogram and a variety of LLDs emotional information. The average accuracy of this model reached 68%, and it greatly compensated for the problem of poor classification effect on the "relaxed" emotion.

4.4. Lyrics Experiment. This group of experiments utilizes different classification models to verify lyrics classification performance. The classification results are shown in Table 3.

Table 3: Accuracy of different lyrics classification models.

Classification models	Angry	Happy	Relaxed	Sad	Average
Word2vec + CNN	0.584	0.62	0.615	0.671	0.622
Word2vec + LSTM	0.685	0.709	0.647	0.731	0.693
Word2vec + CNN-LSTM	0.721	0.823	0.627	0.742	0.728
Our model	0.752	0.815	0.646	0.756	0.742

Table 4: Accuracy of different multimodal fusion methods.

Fusion methods	Angry	Happy	Relaxed	Sad	Average
Feature fusion	0.742	0.712	0.682	0.762	0.724
Decision fusion	0.78	0.765	0.705	0.782	0.748
Our fusion method	0.808	0.823	0.726	0.773	0.782

Table 5: Performance comparison of our model with existing models.

	Modal	Time	Classification	Accuracy
Seo, Huh [7]	Audio	2019	LLDs + SVM	0.571
Zhao et al. [9]	Audio	2018	MIDI + RNN	0.568
Chen, Tang [11]	Lyrics	2018	TF-IDF	0.622
Reddy, Mamidi [14]	Lyrics	2018	Word2vec + LSTM	0.693
Rachman et al. [24]	Multimodal	2018	Random forest feature fusion	0.738
Shi, Feng [25]	Multimodal	2018	LFSM decision fusion	0.758
Su, Xue [26]	Multimodal	2017	Sentence-level decision fusion	0.806
This paper	Multimodal		Multifeature combined classifier + stacking fusion	0.782

Experiments show that LSTM has a better ability to process serialized text data, but CNN has not achieved good performance on text. Meanwhile, the use of CNN-LSTM combined network has obtained a better classification effect, which is improved by 3% compared with the single LSTM method.

Compared with the classification model using only Word2vec features, the multifeature combined network classifier proposed in this paper has achieved the best classification effect. The model integrates Word2vec and chi-squared test vector, and the average classification accuracy reaches 74%.

4.5. Multimodal Fusion Experiment. This group of experiments verifies the classification performance of different multimodal fusion methods. The feature fusion used in the experiment is a concatenated method with normalization, and then, the multifeature combined network classifier is input. The decision fusion is a linear probability voting method (the probability factor is 0.5) to fuse the classifier results. The classification results are shown in Table 4.

The experimental results show that, due to the heterogeneity of different modal features, the classification accuracy of the feature fusion method is not well enough and some emotional information is lost in the process of dimensionality reduction and normalization. The accuracy of the decision fusion method reaches 75%, but the relationship between the modals is ignored with a problem of low scalability at the same time.

The multimodal ensemble method based on stacking proposed in this paper has obtained the best performance, with an accuracy of 78%. Compared with the single-modal classifier, the accuracy has been improved by 4%.

4.6. Comparative Experiment. In order to further verify the effectiveness of the classification model in this paper, the classification models proposed by other researches in the field of music sentiment classification in recent years have been applied in the experimental dataset of this paper for performance comparison. The average accuracy results are shown in Table 5.

The experimental results show that the proposed method has a significant improvement in comparison with the existing popular methods. Compared with merely using audio or lyrics features, multimodal classification performance is better than single-modal classification. Compared with traditional SVM, CNN, and RNN methods, the multifeature combined network classifier proposed in this paper has better classification effects. And, the stacking fusion method also played a role. Su and Xue [26] provided inspiration for the next research. We can continue to split the samples to enhance the classification performance.

5. Conclusions

Identifying specific emotions from music is a challenging task, and the results often depend on the accuracy of the feature and the validity of the model. In our study, the audio dataset was optimized through fine-grained human voice separation preprocessing, and a multifeature combined network classifier based on CNN-LSTM was proposed. The classifier combines heterogeneous 2D and 1D emotional

features and has been effectively used in audio and lyrics classification, with a high classification accuracy. In order to give full play to the emotional representation capabilities of different modal information, we proposed a multimodal ensemble learning method based on stacking. Compared with single-modal classification, it has outstanding classification effect and remarkable generalization ability.

In future research, we will further optimize the network model, adjust the parameter size to enhance the effectiveness of the classifier and expand emotion categories to perform more fine-grained emotion analysis on music as well.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 3, pp. 1–30, 2012.

[2] F. C. Hwang, J. S. Wang, P. C. Chung, and C. F. Yang, "Detecting emotional expression of music with feature selection approach," in *Proceedings of the International Conference on Orange Technologies*, pp. 282–286, IEEE, Tainan, Taiwan, March 2013.

[3] J. L. Zhang, X. L. Huang, L. F. Yang, Y. Xu, and S. T. Sun, "Feature selection and feature learning in arousal dimension of music emotion by using shrinkage methods," *Multimedia Systems*, vol. 23, no. 2, pp. 251–264, 2017.

[4] R. G. Ramani and K. Priya, "Improvised emotion and genre detection for songs through signal processing and genetic algorithm," *Concurrency and Computation: Practice and Experience*, vol. 31, no. 14, Article ID e5065, 2019.

[5] S. H. Chen, Y. S. Lee, W. C. Hsieh, and J. C. Wang, "Music emotion recognition using deep Gaussian process," in *Proceedings of the 2015 asia-pacific signal and information processing association annual summit and conference (APSIPA)*, pp. 495–498, IEEE, Hong Kong, China, December 2015.

[6] Y. C. Lin, Y. H. Yang, and H. H. Chen, "Exploiting online music tags for music emotion classification," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 7, no. 1, pp. 1–16, 2011.

[7] Y.-S. Seo and J.-H. Huh, "Automatic emotion-based music classification for supporting intelligent IoT applications," *Electronics*, vol. 8, no. 2, p. 164, 2019.

[8] W. Yaxin, *Personal Music Emotion Analysis Based on BP Neural Model*, Institute of Management Science and Industrial Engineering, Bangkok, Thailand, 2018.

[9] W. Zhao, Y. Zhou, Y. Tie, and Y. Zhao, "Recurrent neural network for MIDI music emotion classification," in *Proceedings of the 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pp. 2596–2600, IEEE, Chongqing, China, October 2018.

[10] Y. An, S. Sun, and S. Wang, "Naive Bayes classifiers for music emotion classification based on lyrics," in *Proceedings of the 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, pp. 635–638, IEEE, Wuhan, China, May 2017.

[11] X. Chen and T. Y. Tang, "Combining content and sentiment analysis on lyrics for a lightweight emotion-aware Chinese song recommendation system," in *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, pp. 85–89, Macau, China, February2018.

[12] H. He, B. Chen, and J. Guo, "Emotion recognition of pop music based on maximum entropy with priors," in *Pacific-asia Conference on Knowledge Discovery and Data Mining*, Springer, Berlin, Germany, 2009.

[13] C. S. Lee, M. H. Wang, L. C. Chen et al., "Fuzzy semantic agent based on ontology model for Chinese lyrics classification," in *Proceedings of the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 4254–4259, IEEE, Miyazaki, Japan, October 2018.

[14] G. R. R. Reddy and R. Mamidi, "Addition of code mixed features to enhance the sentiment prediction of song lyrics," 2018, https://arxiv.org/abs/1806.03821.

[15] J. Wang and X. Zhao, "Deep Learning Based Mood Tagging for Chinese Song Lyrics," 2019, https://arxiv.org/abs/1906.02135.

[16] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proceedings of the Interspeech 2017*, pp. 1089–1093, Stockholm, Sweden, August 2017.

[17] J. Kim and R. A. Saurous, "Emotion recognition from human speech using temporal information and deep learning," in *Proceedings of the Interspeech 2018*, pp. 937–940, Hyderabad, India, September 2018.

[18] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, 2019.

[19] R. Wang, Z. Li, J. Cao, T. Chen, and L. wang, "Convolutional recurrent neural networks for text classification," in *Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6, IEEE, Budapest, Hungary, July 2019.

[20] J. Zheng and L. Zheng, "A hybrid bidirectional recurrent convolutional neural network attention-based model for text classification," *IEEE Access*, vol. 7, pp. 106673–106685, 2019.

[21] R. Mihalcea and C. Strapparava, "Lyrics, music, and emotions," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 590–599, July2012.

[22] R. Panda, R. Malheiro, B. Rocha et al., "Multi-modal music emotion recognition: a new dataset, methodology and comparative analysis," in *Proceedings of the International Symposium on Computer Music Multidisciplinary Research*, Vancouver, Canada, May 2013.

[23] D. Su, P. Fung, and N. Auguin, "Multimodal music emotion classification using AdaBoost with decision stumps," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3447–3451, IEEE, Vancouver, Canada, May 2013.

[24] F. H. Rachman, R. Sarno, and C. Fatichah, "Music emotion classification based on lyrics-audio using corpus based emotion," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 3, p. 1720, 2018.

[25] W. Shi and S. Feng, "Research on music emotion classification based on lyrics and audio," in *Proceedings of the 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pp. 1154–1159, IEEE, Chongqing, China, October 2018.

[26] F. Su and H. Xue, "Graph-based multimodal music mood classification in discriminative latent space," in *Proceedings of the International Conference on Multimedia Modeling*, Springer, Bangkok, Thailand, pp. 152–163, May2017.

[27] E. Jokinen, R. Saeidi, T. Kinnunen, and P. Alku, "Vocal effort compensation for MFCC feature extraction in a shouted versus normal speaker recognition task," *Computer Speech & Language*, vol. 53, pp. 1–11, 2019.

[28] P. H. Shahana and B. Omman, "Evaluation of features on sentimental analysis," *Procedia Computer Science*, vol. 46, pp. 1585–1592, 2015.

[29] N. Chand, P. Mishra, C. R. Krishna, E. S. Pilli, and M. C. Govil, "A comparative analysis of SVM and its stacking with other classification algorithm for intrusion detection," in *Proceedings of the 2016 International Conference on Advances in Computing, Communication, & Automation (ICACCA)(-Spring)*, pp. 1–6, IEEE, Dehradun, India, April 2016.