ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Research on Path Innovation and Mechanism of Music Dissemination in the Digital Era

Jiang Jiang[1,*]

[1] Xinxiang University Music College, Xinxiang, Henan, 453003, China

SUMMARY: *The coming of the digital revolution has brought the network era, which has caused big changes to the music spread situation and enlarged both the range and degree of music propagation. The present research paper proposes one hybrid music recommendation method which is integrated inside one autoencoder. This content in the basic sense can be divided into two parts: one is the analysis of data content, and the other is experiments about the recommendation method. Firstly, a music feature extraction scheme applicable to music recommendation scenarios is designed; secondly, a music hybrid recommendation model is constructed to realize music track recommendation for different users by incorporating the autoencoder structure. Finally, the Jitterbug popular BGM features are extracted from the time domain and frequency domain in the Jitterbug popular BGM audio dataset as well as in the GTZAN public dataset, respectively, to analyze the distribution law, and the model of this paper is compared with other models in the dataset for comparison experiments. The results show that Jitterbit popular BGMs have fundamental tone periodicity, i.e., they maintain a low frequency and energy overall. And this paper's model has the optimal model efficacy with 9.06% and 4.11% improvement in Recall metrics and 4.67% and 2.28% improvement in F1 metrics on the two datasets, and it also has the optimal anti-sparse ability. The model performs well in the accuracy of music recommendation and can provide technical support for the digital music dissemination path.*

KEYWORDS: *feature extraction; digital dissemination; self-encoder; hybrid recommendation*

1 Introduction

In current digital era, music spread has got new changes in both content and form. Quite a lot of music platforms have seen obvious development, therefore they have become a new road for music to spread out. Traditional music spread ways for example line below concerts, on-spot shows, TV, and broadcast wireless have met difficulty to not fall behind the new situations. Therefore, thus, it is extremely necessary that we conduct exploration on new methods for music communication within the digital age[1-3]. In the current digital era, the channels for music spreading show a many-sided and experience-engaging character. Very many communication platforms are complexly connected together, have overlaps between each other, and are carrying out deeper integration now. This procedure, therefore, greatly enlarges the coverage and the influence of music communication activities. For example, music flowing network platforms such as NetEase Cloud Music and other alike ones have started cooperative works among themselves and with social platforms. This joint work has caused that the channels are combined together. As the consequence, users at present can carry out music

*17516388187@163.com
https://doi.org/10.65102/is2026593 Published: 30 April, 2026

sharing from these music platforms toward many kinds of social media websites. The aim of this sharing is to promote the spreading of music creations. This point manifests a superiority that greatly exceeds that which the traditional music spreading pattern is able to provide [4-6]. Along with the progression of cloud computing, big data, and other technological improvements, the storage abilities of the mobile Internet have seen a rapid promotion. This promotion has caused the big-scale manufacture and extensive spreading of music goods.The maximum number of newly uploaded songs per day on major music websites can reach more than 10,000 songs, of which, more than 2,000 original songs, and the number of cover original accompaniment works is more than 10 million [7-9]. It can be seen that music cultural products have shown the trend of sea quantization. Then, the development of social media provides a new path for music dissemination, which improves the speed and efficiency of the dissemination of music works with the dual perception of hearing and vision [10, 11].

And under the support of digital technology, it helps the music dissemination to be more accurate, comprehensive, effective and safe. Hou [12] used the convolution nerve network method to conduct examination on music data coming from music platform and users' innovation hobby. After that step, a music recommendation system which has individual customization has been completed. This system possesses the capability to provide the platform's users a more accurate music recommendation service, hence promoting the efficiency of music propagation.You [13] predicted the influence of Chinese traditional music in microblog communication in the digital era with the help of a hybrid deep learning model with bidirectional gated recurrent units and dynamic graphical attention networks, which helps to adjust music communication strategies instantly. Su and Sun [14] used deep learning techniques for IoT-based AI image detection for recognizing and classifying the quality of instant music video content in social media and other platforms to provide healthier and higher-quality results for the distribution of musical works. Liu and Zou [15] designed an automatic music score recognition system based on artificial intelligence in a music entertainment environment to improve user interactivity and entertainment experience during the online dissemination of musical works as a way to improve the dissemination effect. Park et al [16] pointed out that the virtual concert solves the problem of insufficient experience in traditional online concerts, subverts the online concert model with immersive experience, creates a new music consumption scene, and provides a new music communication path. Li and his work group [17] utilized the blockchain technology to have developed a decentrally organized music copyright operation and management system. This system carries out the protection for the benefits of many different interested parties, which include creators, copyright holders, operators, and users, all together. It promotes the sustainable development of the music domain and protects the spread of music works. Under this background, to carry out research on creative methods of music transmission in the digital era and their inner transmission mechanisms has great significance for the effect of music spreading.

In this paper, we start from the music content features, introduce the methods in audio processing, and then design a standardized music feature extraction scheme for music recommendation scenarios. On this basis, autoencoder technology is introduced, and for the deficiencies in the traditional autoencoder, noise reduction autoencoder is used to optimize and construct the music hybrid recommendation model. At the same time, the selected datasets are extracted from the time domain and frequency domain respectively to summarize the laws of audio features. Then the model of this paper is compared with the other three comparative models in terms of Recall, Precision, and F1 indexes on different datasets, and the results are analyzed to verify the actual effect of the model in music dissemination recommendation.

2 Characterization of musical content

2.1 Audio Processing Methods

2.1.1 Audio data format

WAV is one kind of audio document form that Microsoft develops, which follows the RIFF (Resource Interchange File Format) document regulation. The content inside the document is constituted by data block pieces. It is employed by people for the storage of audio information on the Windows platform, and it obtains extremely extensive support from application programs, which supports a variety of compression algorithms, support for a variety of audio digital, sampling and It supports a variety of compression algorithms, a variety of audio numbers, sampling and sound channels, the standard format of wav files and the same format as the CD, are 44100Hz sampling frequency, 16-bit quantization of numbers, and the quality of the sound and the CD is comparable to that of the high-quality audio file format.

Because of the wide support and versatility of the wav format, the audio files used in this paper are uniformly transcoded wav files.

2.1.2 Fourier variations and spectra

In the actual computer audio documents that exist in the real world, because what they store are discrete data particulars rather than continuous wave forms, therefore it becomes an essential thing that we must use the discrete Fourier transform for completing the conversion from time domain to frequency domain.For example, for a continuous signal $x(t)$, its corresponding continuous Fourier transform $x'(\omega)$ are continuous functions. In order to discretize $x(t)$ and $x'(\omega)$ and create the corresponding Fourier transforms. If the time region of $x(t)$ is from 0 to L, discretize $x(t)$ in the time domain, which results in a finite-length discrete signal $x_{discrete}(\mathrm{t})$. Let the sampling period be T, then the number of time-domain sampling points is $N = L/T$.

$$x_{discrete}(t) = x(t)\sum_{n=0}^{N-1}\delta(t-nT) = \sum_{n=0}^{N-1}x(nT)\delta(t-nT) \tag{1}$$

Its corresponding discrete Fourier transform is then:

$$x'_{discrete}(\omega) = x(t)\sum_{n=0}^{N-1}x(nT)F\delta(t-nT) = \sum_{n=0}^{N-1}x(nT)e^{-in\omega T} \tag{2}$$

Below presents the continuous Fourier transform which belongs to a function that holds continuous values $x(t)$ sampled in the time domain, i.e., the discrete-time Fourier transform, which is still continuous in the frequency domain.

The next step is to convert the continuous frequency domain signal to a finite length discrete signal as well. Using a similar process for time domain signals, the finite-length discrete signal obtained after discretization assumes that the frequency domain signal is finite. According to the sampling theorem, the frequency domain signal $x'(w)$ has a range of $[0,1/(2T)]$ if the time domain sampling is to be able to completely reconstruct the original signal. Since the range of the time-domain signal is $[0,L]$, according to the sampling theorem as well as the time-frequency dyadic relationship, the sampling interval in the frequency domain can be obtained as $1/L$. According to this, the number of sampling points that lie in the frequency domain is

$$\frac{1/T}{1/L} = N \tag{3}$$

The number of sampling spots inside the instant and frequency fields is also N. The frequency of sampling point positions in the frequency domain is $2\pi k / NT(0 <= k < N)$ on the discrete Fourier transform:

$$x'[k] = x'_{discrete}\left(\omega_k\right) = \frac{1}{T}\sum_{n=0}^{N-1} x[nT]e^{-i\frac{2\pi}{N}nk} \tag{4}$$

In the above equation, by making T 1, the discrete Fourier transform as defined before is obtained. Therefore, the so-called discrete Fourier transform is the result of solving for its corresponding continuous Fourier transform and then discretizing it in the frequency domain.

After doing the discrete Fourier transform on an audio file, the spectrum of that audio is obtained. The so-called spectrum is the distribution in frequency of waves with different amplitudes over a period of time.

2.2 Music Feature Extraction

2.2.1 Feature Extraction Process

Sample Interception: The audio data volume of music resources is usually large, and it is time-consuming to process all of them.

Audio pre-processing: It mainly includes three steps: format conversion, selective filtering of audio data, and spectral transformation.

Extraction of overall features: Calculate the pitch features, change features and rhythm features of the music as a whole, which reflect the corresponding attributes of the music melody.

Local music features: Calculate the Mel frequency cepstrum coefficients of each frame in the sample, and obtain the representative results among them as local features by clustering method.

Integration output: the vectors of overall features and local features obtained before are merged, and finally a sixteen-dimensional vector is obtained as the corresponding music content feature of the audio.

2.2.2 Sample Interception and Preprocessing

In this paper, the audio clips are feature extracted using the later method for the samples to be selected and the main clip, and the Euclidean distance between the features of the samples to be selected and the main clip is calculated separately, and the test finds that more than half of them possessing the highest total energy samples have the highest similarity with the features of the main clip. Therefore, this paper uses the highest total energy of the samples to be selected as the samples for subsequent processing.

2.2.3 Overall Musical Characteristics

This section focuses on feature extraction from the perspective of the overall musical melody. The composition of the melody varies greatly from music to music, and this paper is ultimately for the recommendation to focus on the music style. It is meaningless and computationally intensive to extract the whole melody. Instead of complex melody extraction, this section defines the features of music melody as pitch, variation and rhythm, and extracts the overall features based on audio spectral data.

2.2.4 Local musical characteristics

1. MFCC

Mel Frequency Cepstrum Coefficient (MFCC) is a widely used feature in speech recognition, which can filter irrelevant information such as noise in the audio and reflect the recognizable components of the audio signal. Because the characteristics of music and speech recognition have some similarity, the Mel Frequency Cepstrum Coefficient (MFCC) can also achieve better results in the processing of related music.

2. Gaussian Mixture Model

Gaussian Mixture Model (GMM) is a model for building probability density distribution. The model consists of a mixture of multiple Gaussian models, where each Gaussian model represents a cluster. The sample data is undergone training for ascertaining its probability density distribution on many different categories. This model obtains the widespread application in biometry domains, which include speech processing.

A Gaussian mixture model consisting of M clusters is the weighted sum of the densities of the individual Gaussian models in it, as in Equation (5):

$$p(x \mid \lambda) = \sum_{i=1}^{M} \omega_i g\left(x \mid \mu_i, \sigma_i\right) \tag{5}$$

where x is a multidimensional vector (e.g. of computed features). ω_i are the corresponding weights for each Gaussian model. $\omega_i g\left(x \mid \mu_i, \sigma_i\right)$ is then the density function of each Gaussian model, and the corresponding probability can be obtained based on x as in Eqn. (6) where μ_i is the center vector, σ_i is the covariance matrix, and all the ω_i arithmetic sums to 1.

$$g\left(x \mid \mu_i, \sigma_i\right) = \frac{1}{(2\pi)^{D/2}\left|\sigma_i\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(x-\mu_i\right)' \sigma_i^{-1}\left(x-\mu_i\right)\right\} \tag{6}$$

Thus the complete Gaussian mixture model includes μ_i, σ_i and ω_i for each of these Gaussian models. Given a set of data, when finding the parameters of a Gaussian model, it is usually expected that the output of each data point in the probability density function is maximized, and the computational objective is shown in Equation (7):

$$\max \sum_{i=1}^{N} \log g\left(x \mid \mu_i, \sigma_i\right) \tag{7}$$

3 Autoencoder-based propagation model for music mix recommendation

3.1 Auto Encoder

3.1.1 Self-Encoder Model

Self-encoder, which is a important sub-area in the field of deep learning, is one kind of unsupervised learning method that lets output copy input data. One conventional type of self-encoder is composed by an encoder and one decoder, which are symmetrical in their structure, the encoder is mainly used for the realization of data compression, the encoder maps the data from high-dimensional to low-dimensional, so that the data volume has been reduced by a factor

of geometrical magnitude, The decoder has the function of executing decompression work. When the decoder is placed into working state, it receives the data which the encoder outputs, hence it makes the re-construction of the original input data possible. This one course may be described with one equation:

Decoding Process:

$$h_1 = \sigma_e\left(W_1 x + b_1\right) \tag{8}$$

Coding process:

$$y = \sigma_d\left(W_2 h_1 + b_2\right) \tag{9}$$

where W_i, b_i are weight and bias terms and σ is a nonlinear transformation.

3.1.2 Noise-canceling encoder modeling

The principle of the traditional self-coder to achieve the score prediction lies in the data compression and dimensionality reduction and then restore the dimensionality, in this process, if there is weak correlation between the data in the matrix of correlation data, then in the matrix through the self-coder processing this part of the weak correlation of the data is easier to be destroyed, making it difficult to restore the initial value but transformed to meet the data through the majority of correlation between the strong correlation between the Correlation calculated by most of the strong correlation between the data, so in the training of the autocoder can be strong correlation in the matrix to highlight the correlation information, reduce the impact of outliers, so as to achieve the effect of stabilizing the extraction of features of the matrix, but at the same time, if the training is too many times, or the input data scoring matrix has a strong correlation between the data of the regularity of the data, then the autocoder is just like a complexity of a high degree of the equals sign, the meaning of the training is lost and can not highlight the matrix. It loses the significance of training and fails to highlight the matrix features, at this point, it is urgent to propose a self-encoder that can avoid this situation.

The noise-reduction auto-encoder model, as it is constructed deliberately, has a certain amount of Gaussian noise incorporated by it into the complete input signal. This makes the complete input signal undergo local destruction in different degrees, therefore it produces a signal that has noise and destruction. After that, this signal which has noise and damage is sent by people into the auto-encoder. After passing through the encoding and decoding procedures, the auto-encoder reconstructs an output signal which is the same as the one that is gotten when the complete input signal is put in. In this place, the place at which Gaussian noise is put in corresponds to the position of the input signal, and the position of the added Gaussian noise is the same as the position of the input signal. The locations where Gaussian noise is added and the blank data are replaced by the data predicted by the model, which conforms to the extracted correlation features. The basic structure of the noise-canceling auto-encoder is shown in Fig. 1.

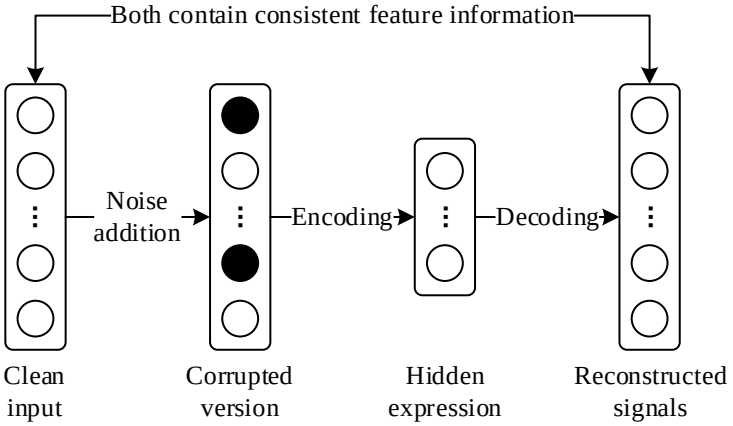


Figure 1: Schematic diagram of noise reduction auto-encoder structure

As with traditional self-encoders, a reconstructed signal that is as similar as possible to the output obtained from the complete signal input can only be obtained when the loss function is minimized.

The emergence of the noise-reducing autocoder changes the situation where the traditional autocoder becomes a meaningless constant function, making the feature representation of the hidden layer units more robust, but due to the artificial noise interference with the complete raw data before each training, the processing time of the model increases dramatically, and has a more serious flaw in the efficiency problem.

3.2 Scoring function

In the cooperative filtration method, when we have the goal to provide a new recommendation for a user, there exist two main kinds of methods. As for the recommendation which is based on users, the process includes finding out the items that users who have similar preferences have given favorable comments to. On another hand, inside the item-based recommendation, the most items which have close relation are calculated. After that step, these project entries are combined by the system to generate the final recommendation result. When we handle the predicted set, we confine our consideration to the binary situation.

In the recommendation which centers on users, the rating function that is used for producing recommendations is shown in the equation (10):

$$h_{ui}^{U} = \sum_{v \in U} f\left(w_{uv}\right) r_{vi} = \sum_{v \in U(i)} f\left(w_{uv}\right) \tag{10}$$

where h_{ui}^{U} denotes the rating of music i by user u, w_{uv} denotes the similarity between user u and user v, and r_{vi} denotes the rating of music i by user v. In the binary case, there are only two values, 0 and 1, that indicate whether user v has listened to music i or not, so it can be simplified to the right-handed equation, $U(i)$ denoting the set of users in the test set who have listened to music i in common with user u. That is, user u's rating h_{ui}^{U} for music i is proportional to the degree of similarity between u and other users v who have listened to music i.

After considering computational efficiency and data materials, this article therefore chooses the item-based recommendation scoring function. This function has connection with the recommendation which is based on user, and the recommendation which is based on item is shown in Equation (11):

$$h_{ui}^{S}=\sum_{j\in L} f\left(w_{ij}\right) r_{uj}=\sum_{j\in L(u)} f\left(w_{ij}\right) \tag{11}$$

where h_{ui}^{S} denotes the rating of music i by user u, and w_{ij} denotes the degree of similarity between music i and music j, which is positively proportional to the degree of similarity between music i and music j that has been listened to by other user u.

The function $f(w)$ is monotonically increasing, this function has the effect of either emphasizing or reducing the importance of the similarity’ s effect on the recommendation. In the actual experiments which we carry out, its influence on the results has very big importance.

Therefore, inside the recommendation modeling work, it is needed that we model the degree of similarity w_{ij} between music i and music j.

The goal of making use of training data is to generate the features which are implicitly expressed by this data for the music $V=\left(V_1 \ldots V_t\right), V_j$ denotes the implied feature vector for the jth music, this text gives the description and depiction of the music in many different aspects. After obtaining the implied features V of the music, the degree of similarity w_{ij} between the music i and the music j is calculated by the cosine similarity as shown in Equation (12):

$$w_{ij}=\frac{V_i \cdot V_j}{\left|V_i\right|\left|V_j\right|} \tag{12}$$

While the generation of music implicit features V is the key to the model, the next section mainly builds a deep learning model from shallow to deep to generate implicit feature vectors for music.

3.3 Feature extraction

3.3.1 Audio Characterization

Audio features we used We use M_j to denote the MFCC feature for the j th music and C_j to denote the Chroma feature for the j th music.

3.3.2 Characterization of lyrics

The TF-IDF measurement for a special word is used to measure its importance inside a certain document which is in a text collection. The importance of the word is directly connected to the speed with which it appears in that particular document, which is called the word frequency (TF). But at the same time, it has opposite connection with the frequency of its occurrence in the whole corpus, this is what people call Inverse Document Frequency (IDF).

Word frequency expresses the speed that a specific word occurs inside a text. In order to avoid the distortion of this frequency which is caused by differences in document length, this value is conducted standardization processing with respect to the total word number in the document, and the formula for the word $\mathrm{t_i}$ and the document d_j, tf_{ij} is shown in Eq. (13):

$$tf_{ij}=\frac{n_{ij}}{\sum_k n_{kj}} \tag{13}$$

In the above equation, n_{ij} denotes the number of times the word t_i occurs in the document d_j, and the denominator $\sum_k n_{kj}$ denotes the total number of all words in the document d_j.

The Inverse Document Frequency (IDF) indicates the importance of the word, idf_i for a word t_i, by dividing the total number of documents by the total number of documents containing the word, and subsequently after taking the quotient in logarithmic form, as shown in Equation (14):

$$idf_i = \log \frac{|D|}{|D:i|+1} \tag{14}$$

$|D|$ denotes the total number of documents in the corpus, and $|D:i|$ denotes the number of documents containing the word, and the use of $|D:i|+1$ in the denominator is to prevent the word from not being in the corpus, which would result in a denominator of zero.

Finally, the $tfidf_{ij}$ value of the word t_i for the document d_j can be expressed as shown in Equation (15):

$$tfidf_{ij} = tf_{ij} \times idf_i \tag{15}$$

From the formula we can see that $tfidf_{ij}$ tends to keep the words that are more important to the document and filter to the common words.

The $tfidf$ value for each word is calculated in the Million Song Dataset dataset. And now, the lyrics of the songs are our documents, for the lyrics d_j, its lyrics feature vector L_j can be represented as shown in Equation (16):

$$L_j = \frac{\sum_i t\,fidf_{ij} \times v_i}{\left|d_j\right|} \tag{16}$$

$\left|d_j\right|$ denotes the number of words in the j lyrics, and the formula shows that the important words in the lyrics have higher weights in the feature vector of that lyric.
The pre-trained corpus is not the same as our training data, so there are cases where words do not exist in the lyrics, but since there are fewer words that do not exist, we ignore the words that do not exist. Finally, we generated feature vectors representing the lyrics for each song.

3.4 Hybrid recommendation models

A hybrid music recommendation model is built in conjunction with an autoencoder.The MFCC and Chroma audio features M_j and C_j belong to 2-dimensional features, but the lyrics feature L_j, the user-side rating vector $S^{(i)}$ and belong to 1-dimensional features.When we process one-dimensional features, we use a full-connected automatic encoder to conduct the work of feature extraction. At the final step, this paper has the construction of a hybrid recommendation model done.

4 Experimental design and analysis

4.1 Data sources

In this paper, we started to collect audio from August 15, 2023, and through the above steps, we completed the construction of the Jieyin popular BGM music library on February 10, 2024, which contains a total of 900 BGMs.Since most of the Jieyin popular BGM audios are tens of seconds, in order to correspond to them, this paper proposes to use the GTZAN public dataset as a comparative analysis dataset for the audio features, which has a total of 1000 songs, containing 10 different genre types, each genre has 100 songs, and each song is about 30 seconds.

4.2 Audio Data Analysis

4.2.1 Time domain characterization

After the waveform of the audio signal has obtained standardization, the value of zero-crossing rate corresponds to the frequency that the signal passes through the zero axis inside one single frame. When the signal crosses the zero axis with higher frequency, therefore the zero-crossing rate will get a higher value, which represents that the amplitude change of the waveform is more frequent, and the energy change is more drastic. Using the zero_crossing_rate function (ZCR) in the librosa library in Python, we can get the short-time zero crossing rate of each frame in the audio. Taking the popular BGM “Xiao Mei Man” of Jieyin as an example, the parameters are set to the frame length of 1947 and the frame shift of 483, and the time sequence of the over-zero rate of “Xiao Mei Man” is obtained as shown in Fig. 2.

The mean value of the over-zero rate of “Xiao Mei Man” is 0.0963, the variance is 0.0027, the quarter-quartile point is 0.0624, and the three-quarter-quartile point is 0.1103. The audio of “Xiao Mei Man” has a total of 29.9 seconds, and it can be found in the figure that this 29.9-second audio has a strong periodicity, and a higher over-zero rate is always accompanied by several lower over-zero rates in each cycle. Each cycle is always accompanied by a higher zero crossing rate and a number of lower zero crossing rates. This indicates that the energy of this audio has a periodicity, and the high and low frequencies are frequently transformed.

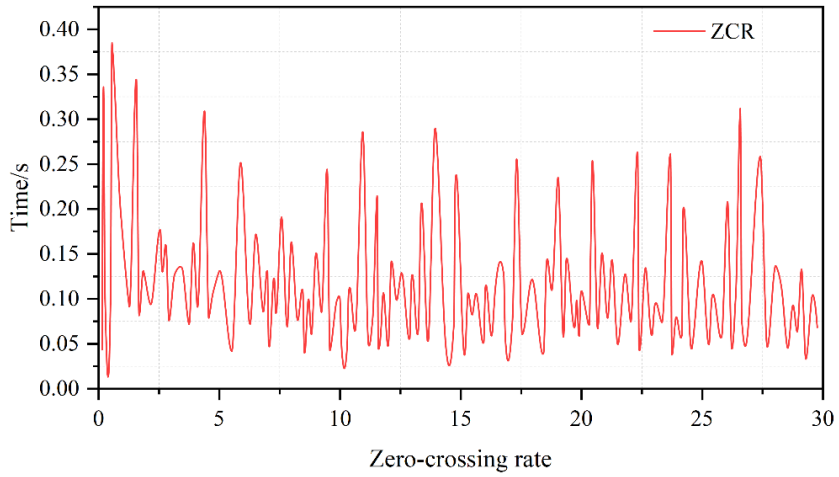
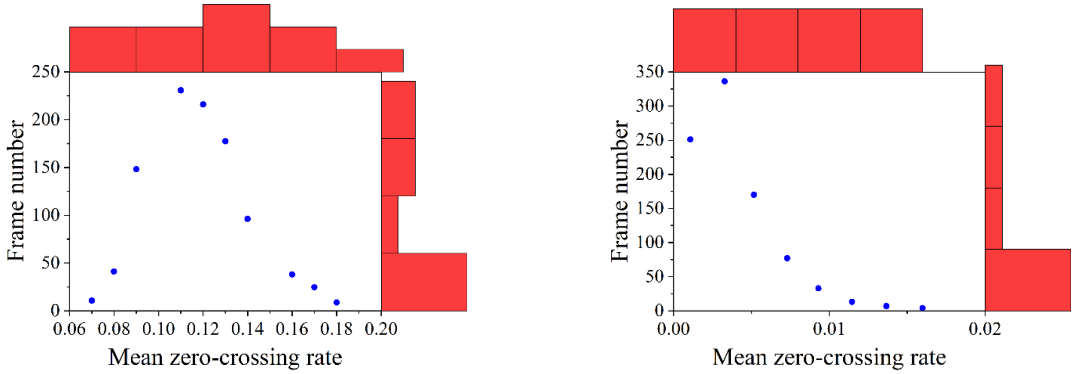


Figure 2: The song "Little Happiness" has a zero pass rate

The mean value of over-zero rate is obtained by summing and averaging the over-zero rate of each frame of the audio, and the distribution of the mean and variance of over-zero rate of

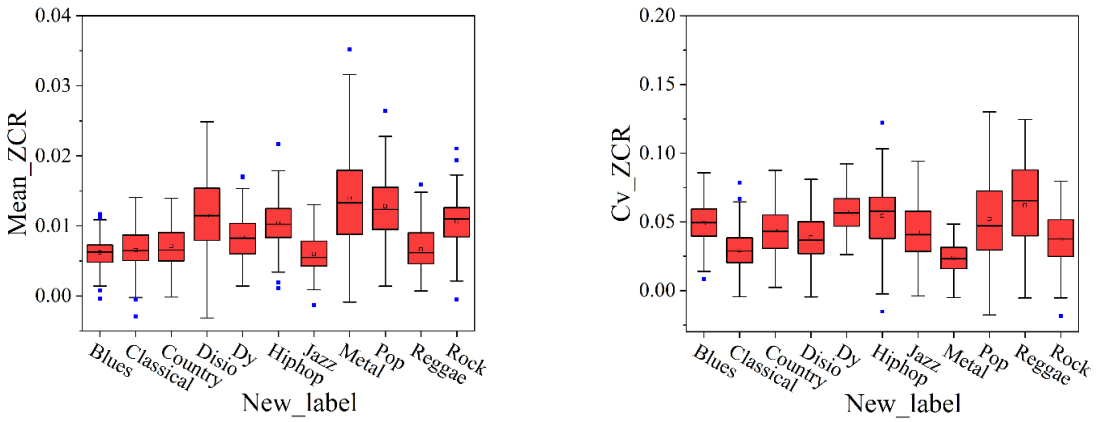
the popular BGMs in Jitterbites is shown in Fig. 3, with (a) and (b) being the distribution of over-zero rate mean and over-zero rate variance, respectively. It can be found that the mean value of the over-zero rate of these 900 BGMs is mainly concentrated in the range of 0.07 to 0.12, which maintains a low over-zero rate interval, which also proves that most of the BGMs in the Shake Music list are turbid sounds with human voices. Observing Fig. 3(b), it can be found that the histogram has a right-skewed distribution, and the variance of the over-zero rate of the majority of the audio is small, indicating that the over-zero rate of each frame within each BGM is very small in the degree of dispersion, just like the song "Xiaomeiman," the majority of the frames keep low over-zero rates, and individual frames have high over-zero rates, and the distribution exists in the periodicality.



(a) Mean zero-crossing rate (b) Zero-crossing rate variance

Figure 3: Mean and variance distribution of TikTok's popular BGM zero-pass rate

We will carry out one comparative analysis by making use of the ten genre types that exist inside the publicly open GTZAN dataset. For the prevention of an obvious difference in the measurement dimension of the short-time zero-crossing rate among different types of music, the variation coefficient is utilized to measure the scattering degree of the short-time zero-crossing rate between frames inside each piece of the audio segment. Figure 4 shows the box line plot of the distribution of the mean value of the short-time over-zero rate and the coefficient of variation among different genres, and it can be found that the short-time over-zero rate of the Jitterbug popular BGMs is at an intermediate level compared with the other 10 genres, whereas the mean value of the coefficient of variation is higher compared with the other 10 genres.



(a) Mean zero-crossing rate (b) Coefficient of variation

Figure 4: Distribution Box Plot

The results of the multi-sample Kruskal-Wallis H test are put in Table 1. Under the significance level 0.05, there is a statistics-wise obvious difference in the distribution of the mean short-term zero-crossing rate among different categories (H = 503.122, p = 0.000), and the results of Bonferroni's multiple comparisons show that all ten of the Jitterbug's popular BGMs and the GTZAN dataset of the genres ($p<0.05$); in the coefficient of variation dispersion of short-time positive interest rates across different categories, there exists a statistically significant difference (H = 645.196, p = 0.000)., and Bonferroni multiple comparison results show that Jitterbug's popular BGM is only non-significantly different from reggae, hiphop, and pop genres ($p>0.05$), the There are significant differences with all other seven genres ($p<0.05$).

Table 1: Multiple Comparison Results

Sample	The mean distribution of short-term zero-crossing rate is identical		The distribution of the coefficient of variation of the short-time zero-crossing rate is the same	
	Test statistic	P value	Test statistic	P value
Dy-jazz	373.154	0.000***	569.093	0.000***
Dy-classical	-331.572	0.000***	-907.042	0.000***
Dy-blues	-265.468	0.000***	-286.275	0.000***
Dy-country	-214.001	0.000***	-445.781	0.000***
Dy-reggae	130.095	0.039**	-85.397	0.175
Dy-hiphop	-244.276	0.000***	-24.783	0.722
Dy-rock	-255.746	0.000***	559.312	0.000***
Dy-disco	577.084	0.000***	-584.243	0.000***
Dy-pop	-584.083	0.000***	109.816	0.081*
Dy-metal	-746.082	0.000***	996.535	0.000***

Note: “*”, “**”, “***” indicate significant at 10%, 5%, 1% level of significance respectively

In summary, the internal frames of the Shake Hot BGM maintain a low short-time over-zero rate, mostly turbid sounds with vocals. Overall, the mean and coefficient of variation of short-time over-zero rate of Jitterbug popular BGM are significantly different from the distribution of other genres.

4.2.2 Frequency domain characterization

The spectrum central point, also called the spectrum first-order distance, displays different numerical values according to the distribution of spectrum energy. When the great part of the spectral energy gathers itself in the low-frequency scope, the numerical value of the spectral centroid is comparatively small. On the opposite side, when the majority of the spectrum energy is gathered in the high-frequency scope, the numerical value of the spectrum centroid is relatively big.Using the spectral_centroid function of the librosa library in Python, we can get the spectral center of gravity of each frame in the audio, which can be merged with the waveform of the audio in one graph after 0-1 normalization. Taking the popular BGM “Xiao Mei Man” of Jieyin as an example, the parameters are set to frame length 2031 and frame shift 501, and the spectral center of mass of “Xiao Mei Man” is obtained as shown in Figure 5.

With regard to the audio work that names "Xiao Mei Man", the average numerical value of the spectral centroid is 0.3126. The square of standard deviation, namely variance, has a value of 0.0233. To the spectral centroid data, the first quartile is 0.2051, thus the third quartile is 0.3684. According to these data, thus it can be seen that most of the spectral centroid numerical values of this audio are located in the range from 0.2 to 0.4, and a small portion of it is concentrated in the range of 0.75 or so and the distribution is more regular, with a periodical nature, which means that the audio of “Xiao Mei Man” has more low-frequency components,

and its frequency shift is 501, which is the same as that of the popular BGM “Little Beauty”. This indicates that the audio of “Little Beauty” has more low-frequency components and its high-frequency part has periodicity.

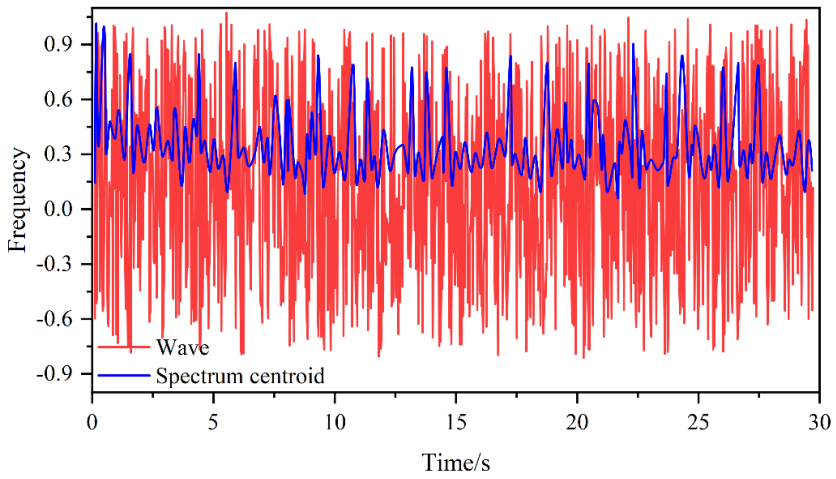
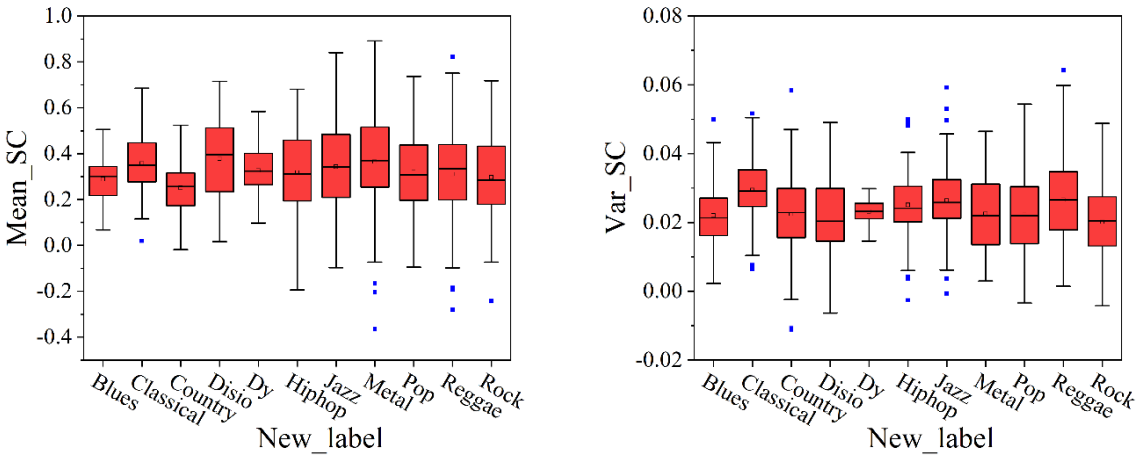


Figure 5: Spectrum centroid diagram of "Little Happiness"

Figure 6 shows the box line plots of the distribution of the mean value of the spectral center of mass and the coefficient of variation among different genres, respectively, and it can be found that the spectral center of mass of Jitterbug's popular BGMs is low compared with the other genres, while the mean value of the coefficient of variation is at an intermediate level compared with the other 10 genres, which suggests that the low-frequency components of each segment of Jitterbug's popular BGMs are more frequent compared with the other 10 genres, and that there is not a large degree of internal discretization.



(a) Mean zero-crossing rate (b) Coefficient of variation

Figure 6: Distribution Box Plot

By utilizing data got from the multi-sample Kruskal-Wallis H test, the results are shown in Table 2. Under a significance level of 0.05, there is a statistics-based significant difference in the distribution of the mean value of the spectral mass center across different categories. The statistics which we use for the test is 152.084, and the p-value is 0.000, the Bonferroni multiple comparisons results show that Jitterbug's popular BGMs have no significant difference only with reggae and rock genres are not significantly different ($p>0.05$), and are significantly different ($p<0.05$) from all other eight genres;on the basis of statistics, a conspicuous difference

may be found in the distribution of variation coefficients for the mass centers of spectrums among different types of works (H = 162.533, p = 0.000), and the results of Bonferroni's multiple comparisons show that Jitterbug's popular BGM is only associated with the jazz, hiphop, rock genres with no significant difference (p>0.05), and with all other seven genres with significant difference (p<0.05).

Table 2: Multiple Comparison Results

Sample	The mean distribution of spectral centroid is identical		The distribution of the coefficient of variation of the spectrum centroid is the same	
	Test statistic	P value	Test statistic	P value
Dy-jazz	-167.342	0.009***	-26.862	0.692
Dy-classical	258.076	0.000***	-906.184	0.000***
Dy-blues	-163.215	0.011***	194.251	0.003***
Dy-country	-316.511	0.000***	290.327	0.000***
Dy-reggae	27.683	0.696	-238.416	0.000***
Dy-hiphop	-234.561	0.000***	106.781	0.091*
Dy-rock	86.579	0.177	-54.308	0.403
Dy-disco	211.836	0.001**	-285.834	0.000***
Dy-pop	-169.06	0.008**	131.439	0.039**
Dy-metal	-433.2	0.000***	479.254	0.000***

Note: “*”, “**”, “***” indicate significant at 10%, 5%, 1% level of significance respectively

To sum up, the broadly liked background music (BGM) of Jitterbug contain a bigger percentage of low-frequency composition parts. The majority of frames maintain a low value of spectral centroid, even though there exist individual frames which display a high value of spectral centroid, and there is a periodicity in the distribution. Overall, the mean values and coefficients of variation of the spectral center of mass of Jitterbug's popular BGM are significantly different from the distributions of other genres.

4.3 Comparison and analysis of experimental results

In order to validate the performance of the hybrid recommendation model algorithms in this paper, the following algorithms are selected for comparison: user-based CF, probability matrix factorization PMF, and collaborative depth model CDL.All the experiments are tested on two datasets with different sparsity levels, TP-100 and TP-500. When we make comparison of the experimental results, two aspects are gotten into consideration. Firstly, the effect degree of the arithmetic methods put forward by this paper is compared with that of the contrasted arithmetic methods under different amounts of recommended music. Secondly, the change of the model's effect is inspected on datasets which have different sparse levels.

4.3.1 Experimental parameterization

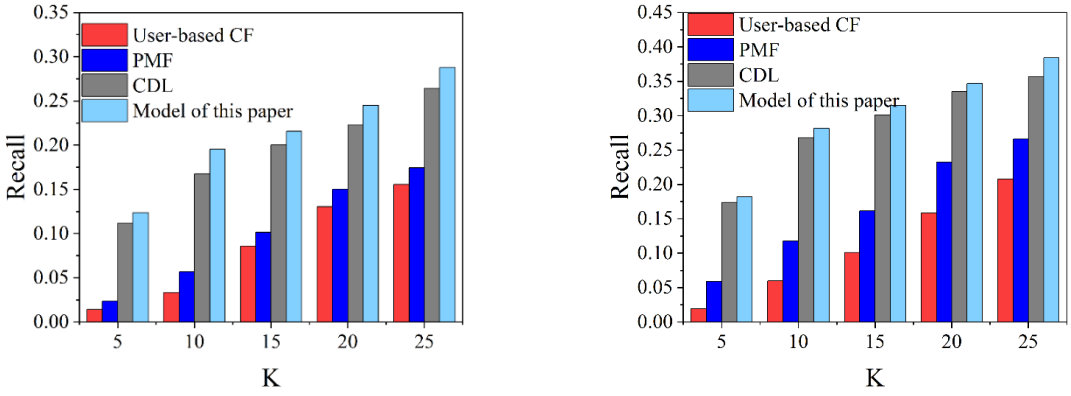
In order to compare fairly, this paper sets the parameters of the comparison algorithms according to the references or experimental results of each comparison algorithm, and the parameter adjustment test method is grid search.The number of neighbors selected for the User-based CF model is 10, and the similarity computation method is Pearson's similarity.The regularization coefficients of the user's hidden feature matrix of the PMF model λ_u = 0.01, item hidden feature matrix regularization coefficient λ_v = 0.5, and hidden feature vector dimension k = 50. In order to have more credibility for the cross-sectional comparison

between the algorithms, the CDL model has the same parameter settings as those shared by the DRCDM model: the regularization coefficient of weight matrix W and bias vector b λ_w=0.01, the regularization coefficient of SAE reconstruction error λ_n=0.1, the number of network layers of SAE L=3, and the number of neuron nodes is set to 4000 → 500 → 50 ← 500 ← 4000, i.e., the number of nodes of the intermediate hidden layer is 50, which is in line with the matrix decomposition of the hidden feature vector dimension is consistent. Hybrid recommendation model hidden layer output and item hidden feature matrix reconstruction error regularization coefficient λ_v=0.5, user and item hidden feature matrix regularization coefficient λ=0.01. Optimization algorithms all use stochastic gradient descent, the learning rate of η=0.01, the maximum number of iterations is set at 100 times, furthermore, the minimal change value of the loss function has been set as 0.01.

4.3.2 Comparative Experimental Results and Analysis

In order to verify the recommendation efficacy of the hybrid recommendation model proposed in this paper, this section conducts experiments on two different sparsity levels of data and respectively in the previous section and compares them with three contrasting algorithms, which are optimally efficacious when cross-comparison is made between the algorithms. Comparisons are made in terms of Recall, Precision for all users, and the results of each experiment are averaged over five-fold cross-validation.

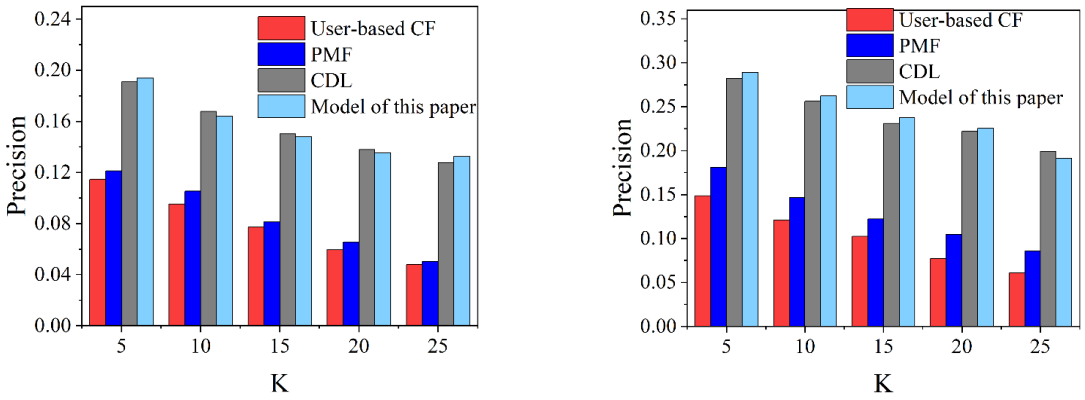
Figure 7 has given a comparison of four models which bases on Recall metrics on two datasets. Through checking the graph, we can clearly see that, among the four models, the mixed recommendation model put forward by this paper achieves the most useful recommendation effect, followed by CDL, PMF, and User-based CF. the Recall values of all the four models show an increasing trend as the number of recommendations, K, grows. This is because the larger the number of recommendations, the larger the number of user-favorite songs covered, and the denominator in the Recall formula is the number of user-favorite songs, which is a constant value. The introduction of content features through deep learning in traditional recommendation algorithms can be seen to significantly improve the effectiveness of the model in the Recall metric. Among the two deep collaborative models, the hybrid recommendation model in this paper is more effective than CDL, and DRCDM in the GTZAN public dataset is 9.06% higher than CDL on average, and DRCDM in the Jitterbit's popular BGM audio dataset is 4.11% higher than CDL on average, which tells us that DRCDM has a better anti-sparse ability under the Recall metric.



(a) GTZAN public dataset (b) TikTok popular BGM audio

Figure 7: The average recall values of the four algorithms under different k values

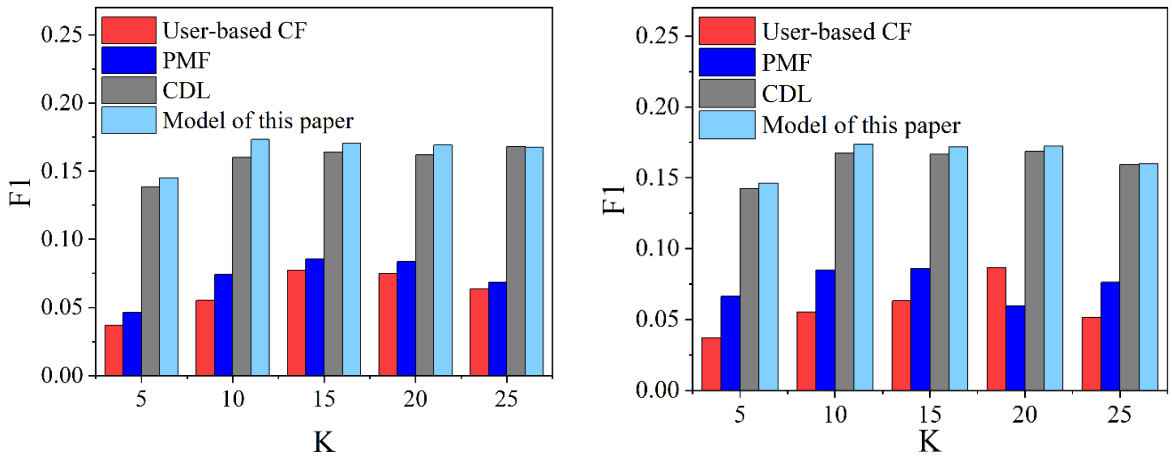
Figure 8 shows the comparison of the four models in terms of precision Precision metrics on the two datasets.According to what the figure shows, when we make comparison among the four models, the hybrid recommendation model and the CDL model have obtained the most good recommendation results. The next that follow are PMF and CF based on user. Along with the quantity of recommendation items goes up, the Precision values of all four models have shown a tendency of going downward.Because of the sparse rating data, the numerator of the Precision metric does not grow as fast as the denominator. The introduction of content features through deep learning in traditional recommendation algorithms can be seen to significantly improve the effectiveness of the model in the Precision metric. Among the two deep collaborative models, the hybrid recommendation model is comparable to CDL, the hybrid recommendation model in the GTZAN public dataset has an average decrease of 0.14% compared to CDL, and the hybrid recommendation model in the Jieyin popular BGM audio dataset has an average increase of 2.17% compared to CDL. And both datasets are when the K value is greater than 20, the effectiveness of the hybrid recommendation model in this paper decreases more than that of CDL.



(a) GTZAN public dataset (b) TikTok popular BGM audio

Figure 8: The average precision values of the four algorithms under different k values

Figure 9 shows the comparison of the four models on F1 metrics in the two datasets. From the figure, it can be seen that among the four models, this paper's hybrid recommendation model achieves the optimal recommendation performance, followed by CDL, PMF, and User-basedCF. When we carry out comparison between the two models, the effect of the hybrid recommendation model which is put forward by this paper is better than that of the CDL model. In the GTZAN public dataset, the hybrid recommendation model has an average promotion of 4.67% when it is compared with the CDL model. In the same way, on the Jitterbit widely used BGM data set, the mixed recommendation model possesses an average promotion of 2.28% when compared with the CDL model. In average situation, the CDL has a decrease of 2.28 percent. This shows that the mixed suggestion model put forward by this paper displays better anti-sparseness ability when it is evaluated through the F1 measurement norms. Along with the increment of the recommendation quantity K, the F1 values of all four models show a trend of increasing and then decreasing. This is the result of the joint action of Recall and Precision indicators, and the increase of F1 will inevitably lead to the increase of coverage and the decrease of precision. In the stage where K is less than 15, the increase of Recall occupies the main part; in the stage where K is less than 15, the decrease of Precision occupies the main part.



(a) GTZAN public dataset (b) TikTok popular BGM audio

Figure 9: The average F1 score of the four algorithms under different k values

Among the two deep models, this paper's hybrid recommendation model has better anti-sparse ability than CDL, and this paper's hybrid recommendation model has better model performance than CDL under Recall and F1 metrics. A side-by-side comparison of the four models reveals that PMF has the worst sparsity resistance, and the effectiveness of PMF decreases dramatically under sparse conditions, even close to User-based CF, and if the content features are introduced into PMF through deep learning to form a hybrid recommendation algorithm, it can not only significantly improve the model's effectiveness under each index, but also improve the model's sparsity resistance.

5 Conclusion

In this paper, we extracted the features that can reflect the music content by processing the audio data, and constructed the music mix recommendation model based on autoencoder by fusion using autoencoder. The Jitterbug popular BGM audio dataset from 2023-8-15 to 2024-2-10 and the GTZAN public dataset are selected, and the time-domain features and frequency-domain features of Jitterbug popular BGM audio are mined by the overshort time-over-zero rate and spectral center of mass of the Jitterbug popular BGM audio, and the comparison experiments are conducted on the two datasets. It is found that there exists a fundamental tone periodicity pattern in Jitterbug popular BGMs, and the hybrid recommendation model proposed in this paper has the optimal model efficacy under Recall metrics and F1 metrics, and it also has the optimal sparsity resistance.

References

[1] Nistor, R. L., & Nedelcut, A. C. (2017). Evaluating the promotion of two music events. Bulletin of the Transilvania University of Braşov. Series VIII: Performing Arts, 73-84.

[2] Pushmin, A. (2023). Music Management: Production System and Promotion in the Music Industry. Socio-Cultural Management Journal, 6(1), 140-164.

[3] Qin, T., & Álvarez, I. C. (2025). Reaching Beyond Traditional Fans: A Study of Early-Music Dissemination, Festivals and Audience Participation. De musica disserenda, 21(2).

[4] Li, X., & Dong, J. (2021, June). User Demand Awareness and Analysis of Online Music——Take NetEase Cloud Music Platform as an Example. In Proceedings of the 2021 5th International Conference on E-Education, E-Business and E-Technology (pp. 82-88).

[5] Jidong, L. (2023). Music Communication Strategies in the Era of Digital Culture Industry: An Examination Centered on QQ Music Platform. Media and Communication Research, 4(7), 35-41.

[6] Qu, S., Hesmondhalgh, D., & Xiao, J. (2023). Music streaming platforms and self-releasing musicians: the case of China. Information, communication & society, 26(4), 699-715.

[7] Zhang, Y. (2025). Characteristics and Reflections on Music Communication from the Perspective of Network Empowerment. Lecture Notes in Education, Arts, Management and Social Science, 3(6), 315-320.

[8] Matuszewski, B. (2020). A web-based framework for distributed music system research and creation. AES-Journal of the Audio Engineering Society Audio-Accoustics-Application.

[9] Kusumawati, R. D., Oswari, T., Yusnitasari, T., Dutt, H., & Shukla, V. K. (2020). Investigating customer satisfaction towards music website in Indonesia and India: a comparative study. International Journal of Digital Signals and Smart Systems, 4(1-3), 17-39.

[10] Lyu, Y. (2022, December). Visualization Communication of Ethnic Music on Short Video Platforms A Case Study of the Mongolian Music on Douyin. In 2022 5th International Conference on Humanities Education and Social Sciences (ICHESS 2022) (pp. 363-373). Atlantis Press.

[11] Ji, Y., Yang, Q., Yang, X., Sun, Y., & Zeng, Z. (2025, August). The transmission path and technological innovation of traditional bamboo flute music in the digital age. In Proceedings of the 2025 International Conference on Generative AI and Digital Media Arts (pp. 1-4).

[12] Hou, R. (2024). Music content personalized recommendation system based on a convolutional neural network. Soft Computing, 28(2), 1785-1802.

[13] You, J. (2024). The Influence of Digitization on the Dissemination of Traditional Chinese Music and Weibo Content Propagation Under Deep Learning. IEEE Access, 12, 13870-13877.

[14] Su, Y., & Sun, W. (2023). Classification and interaction of new media instant music video based on deep learning under the background of artificial intelligence. The Journal of Supercomputing, 79(1), 214-242.

[15] Liu, Y., & Zou, Y. (2025). Application of artificial intelligence based on pattern recognition in music entertainment environment and automatic music recognition. Entertainment Computing, 52, 100848.

[16] Park, J., Choi, Y., & Lee, K. M. (2024). Research trends in virtual reality music concert technology: A systematic literature review. IEEE Transactions on Visualization and Computer Graphics, 30(5), 2195-2205.

[17] Li, Y., Wei, J., Yuan, J., Xu, Q., & He, C. (2021). A decentralized music copyright operation management system based on blockchain technology. Procedia Computer Science, 187, 458-463.