ITC 2/54
Information Technology and Control
Vol. 54 / No. 2/ 2025
pp. 520-535
DOI 10.5755/j01.itc.54.2.36737

Neural Networks and Ensemble Model to Automatic Music Coordination: A Performance Comparison

Received 2024/03/21 | Accepted after revision 2024/06/24

HOW TO CITE: Wang, L. (2025). Neural Networks and Ensemble Model to Automatic Music Coordination: A Performance Comparison. *Information Technology and Control*, 54(2), 520-535. https://doi.org/10.5755/j01.itc.54.2.36737

Neural Networks and Ensemble Model to Automatic Music Coordination: A Performance Comparison

Lu Wang
School of Preschool and Art Education, Xinyang Vocational and Technical College, Xinyang 464000, Henan, China

Corresponding author: luwang11lu@163.com

In order to solve the problems of low classification accuracy, poor quality of generated music, and insufficient consideration of the order and duration of notes in music coordination, this paper adopts a long short-term memory network (LSTM) and ensemble model based on the combination of timing and self-attention mechanism. The experimental model uses the LSTM network to automatically learn the important features of notes, and introduces the timing and self-attention mechanism to enhance the model's ability to pay attention to the note sequence and features, and better capture the long-distance dependencies and emotional changes in music. Compared with the traditional model, the model used in this paper is more detailed in considering the order and duration of notes, and combines emotional labels with audio data to improve the quality of music generation. The experiment is verified by the three music datasets of Lim, Rhyu and Lee. The ensemble model combined with LSTM and self-attention mechanism in this paper performs well in comprehensive evaluation scores and chord classification accuracy, which is significantly improved compared with the traditional LSTM model. The novelty lies in the better integration of the timing relationship and emotional information of the note sequence, which improves the performance of music coordination. The model in this paper achieved 43 points (out of 50 points) and 95.6% in comprehensive evaluation score and chord classification accuracy, respectively. The chord classification accuracy was significantly improved by 3.3% compared with LSTM. It also has unique advantages in model structure design and feature integration, especially in the introduction of timing and self-attention mechanisms, and the combination of emotional labels. It has achieved better results and brought new ideas and methods to the field of music generation.

KEYWORDS: Neural Networks, Musical Coordination, Musical Ensembles, Timing and Self-Attention Mechanism, Musical Emotions

1. Introduction

As the social information technology develops rapidly, various kinds of music ensemble performances begin to emerge in public places, among which the rock music ensemble performance is the most prominent. Studying the coordination and note classification of music ensemble performance can provide a basis for musicians to improve the melody, and also enable listeners to experience the beauty of ensemble music more deeply. At present, the coordination of music ensemble performance and the classification of notes do not fully consider the order and duration of notes, resulting in poor coordination of notes at different time steps, and the characteristics of notes cannot be fully learned automatically, and the accuracy of note classification is not high. Especially for the fusion of musical note emotion changes, it is difficult for the model to coordinate such emotion, resulting in the generated music emotion color is not bright, affecting the quality of music. Accurate and reasonable coordination of all kinds of ensemble music can improve the player's control of rhythm and enhance the audience's satisfaction of music.

With the increasing attention paid to cultural undertakings and the development of neural network models, ensemble music coordination has become a research hotspot and has achieved remarkable results. Clayton et al. proposed an Interpersonal Music Entrainment (IME) model to enhance the understanding and note coordination of IME in different music cultures by integrating culturally shared knowledge and emotions [3]. Chakraborty et al. proposed a model based on recurrent neural network (RNN) to achieve synchronization of instrument control and sensing parameters through mapping mode [1]. Experimental results show that this model can achieve human-machine collaborative performance well [1]. Ye et al. proposed a collection of machine learning models aimed at improving the accuracy of music emotion classification. Experiments have shown that this method can improve the classification accuracy [22]. Medina et al. used multilayer perceptron (MLP) to classify music emotions. By comparing with support vector machine (SVM) and random forest model, MLP obtained an average F measure of 50% in four-quadrant classification, and the prediction value reached 73% [14]. Wood et al. studied the training changes of professional string quartets and recorded the movement data of musicians [20]. The experimental results showed that the similarity of the variable group reached 0.72 [20]. Leman et al. used the Bayesian algorithm to predict timing data in music ensembles. Experiments have shown that this method can solve timing problems to a certain extent [10]. Ray et al. used collective efficacy beliefs to measure the quality of generated music through five groups of experiments [15]. The experimental results showed that collective functional beliefs are suitable for string room ensembles, and the quality of generated music is higher, with a correlation of 0.82 [15]. It can be seen from the above literature that the methods in the above literature have certain improvements in the coordination between music melodies and the accuracy of the classification model after introducing emotion. However, the coordination between music melodies is not high and it lacks emotional generative music. Still a current reality.

To meet social needs and the development of national cultural undertakings, researchers have used neural networks to improve classification accuracy and temporal coordination, promoting the development of scientific research in the cultural field. Gunawan and others used long short-term memory (LSTM) and gated recurrent unit (GRU) networks to improve the accuracy of music note classification [8]. Experiments show that the classification accuracy of the double-stacked GRU model reaches 70%, and the music quality score is 6.85 points (out of 10 points) [8]. Xiao et al. proposed a two-stage attention convolution LSTM network model based on multivariate time series (MTS) prediction [21]. The convolutional layer extracts spatial correlation and LSTM extracts temporal correlation. Experimental results show that the model effectively achieves time series prediction [21]. Khare et al. used convolutional neural networks (CNN) to identify and classify emotions, and used smooth pseudo-Wigner-Ville distribution to convert time-frequency signals into images [9]. The experimental results showed that the lowest accuracy rate reached 91.91% [9]. Yeh et al. evaluated the generation of chord sequences as melody and harmony accompaniment, and the deep learning method performed best in automatic melody coordination [23]. Yin et

al. used deep learning and Markov models to evaluate the coordination of automatically generated music, and the results showed that these two methods are better than traditional models [24]. Fu et al. proposed CNN and LSTM time series prediction models based on the temporal self-attention mechanism [6]. Experimental results prove the effectiveness of this mechanism and achieve the best short-term prediction performance [5, 6, 16]. It can be seen that the LSTM recurrent neural network model is feasible for automatic music coordination, but it fails to fully consider the order and duration of note playing, and its ability to generate musical emotion expression is still insufficient. Therefore, based on the above literature, this paper uses The LSTM recurrent neural network and ensemble model that combines time series and self-attention mechanisms can solve this problem.

This study aims to propose a novel music generation model, which uses a long short-term memory network (LSTM) and an ensemble model that combines timing and self-attention mechanisms to achieve the following goals: improve the classification accuracy of music generation and improve the quality of generated music. This paper conducts collaborative analysis of music melody based on the LSTM network that integrates the temporal attention mechanism and the self-attention mechanism. The experiment preprocessed static vectors to capture duration and emotion features by normalizing pitch, note duration, and duration, and encoding emotion and duration labels. The introduction of temporal attention and self-attention mechanisms enhances the model's automatic attention ability to note sequences and captures long-distance dependencies and emotional changes. The performance of this model was compared with five other neural network models on three music data sets. The results showed that the LSTM and sequence self-attention mechanism models performed significantly better on parameters such as chord classification accuracy, F1 value and evaluation score. Advantages, improved classification accuracy and collaboration.

2 NN Model

2.1 LSTM RNN

LSTM is a special recurrent network model [2, 25]. The forgetting threshold is displayed in Formula (1). Li, Li and other scholars enhanced their ability to recognize action timing and understand context in videos by introducing time attention mechanism [11, 12]. Wei and other scholars enhanced modulation recognition of radar signals by introducing SAM [11, 12]. In addition, the temporal attention mechanism and SAM are introduced to enhance the ability of LSTM model to capture the long-distance dependence and the emotional change of note sequences.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f), \quad (1)$$

where f_t is the activation vector of the forgetting threshold; σ represents the sigmoid function, W represents the weight matrix.

The input threshold formula, displayed in Formulas (2)-(3).

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (2)$$

$$\tilde{c} = tanh(W_c[h_{t-1}, x_t] + b_c), \quad (3)$$

where i_t represents the activation function of the input threshold, and $\tilde{C}_t$ represents the current input cell state.

The element state is specifically expressed in the Formula (4).

$$C_t = f_t C_{t-1} + i_t \tilde{C}_t, \quad (4)$$

where C_t represents the neuron cell state vector.

The output threshold displayed in Formulas (5)-(6).

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t * tanh(C), \quad (6)$$

where o_t represents the activation vector of the output threshold.

2.2 LS-ESNs Model

In order to enhance the ability of echo state networks to simulate multi-scale time features, this paper adopts LS-ESNs to conduct experiments, which are composed of multiple independent reservoirs [4, 27]. The ability of each reservoir to capture inter-

dependence on several time scales is crucial for the autonomous coordination of music. Short-term reservoirs capture short-term dependent features by dependence range m, while long-term reservoirs do so by bypassing connections. The various time-scale echo states that are collected from the three reservoirs represent the original musical sequence, which is then transmitted to the output layer by the echo states in succession.

The LS-ESNs model is divided into three different time-dependent reservoir components, long and short-term reservoirs, and typical reservoirs.

Long-Term Reservoir Status

In long-term reservoirs, $x_{long}(t) \in R^{N\times 1}$ and N represent the size of the reservoir, and the calculation formula is displayed in Formula (7).

$$x_{long}(t)=\gamma\cdot tanh\left(W_{in}u(t)+W_{res}x_{long}(t-k)+(1-\gamma)\cdot x_{long}(t-k)\right), \quad (7)$$

where k represents the length of the skipped step; γ represents the permeability.

Typical Reservoir State

The typical reservoir state is $x_{typical}(t)\in R^{N\times 1}$, and the calculation formula is displayed in Formula (8), where $x_{typical}(t-1)$ represents the state at time $t-1$.

$$x_{typical}(t)=\gamma\cdot tanh\left(W_{in}u(t)+W_{res}x_{typical}(t-1)+(1-\gamma)\cdot x_{typical}(t-1)\right). \quad (8)$$

Short-term reservoir status

Short-term reservoir state $x_{short}(t)\in R^{N\times 1}$ is mainly used to obtain short-term correlation. The calculation formula is as follows formulas (9) and (10), where m represents short-term dependence range and captures correlation characteristics through historical state [13].

$$x_{short}(t)=\gamma\cdot tanh\left(W_{in}u(t)+W_{res}x(t-1)+(1-\gamma)\cdot x(t-1)\right) \quad (9)$$

$$x(t-m+1)=\gamma\cdot tanh\left(W_{in}u(t-m+1)+W_{res}x(t-m)+(1-\gamma)\cdot x(t-m)\right) \quad (10)$$

For each time step t, the expressions representation of the original time series are expressed as $x_{long}(t)$, $x_{typical}(t)$, $x_{short}(t)$ on different time scales by Formulas (7)-(9), and then multi-scale representation is performed by $X(t)=[x_{long}(t),\ x_{typical}(t),\ x_{short}(t)]\in R^{3N\times 1}$. The linear output layer formula is displayed in the following Formula (11).

$$y(t+1)=f_{out}\left(W_{out}X(t)\right). \quad (11)$$

The calculation Formula (11) is reconstructed in matrix form, and f_{out} is set as the identification function, as displayed in the following Formula (12). In addition, the weight of the output layer is trained by minimizing the loss function through the following Formula (13).

$$Y=W_{out}X \quad (12)$$

$$L(W_{out})=||T-W_{out}X||_2^2+\lambda||W_{out}||_2^2. \quad (13)$$

The calculation formula of the weight of the output layer is displayed in Formula (14) below.

$$W_{out}=(X^TX+\lambda I)^{-1}X^TT. \quad (14)$$

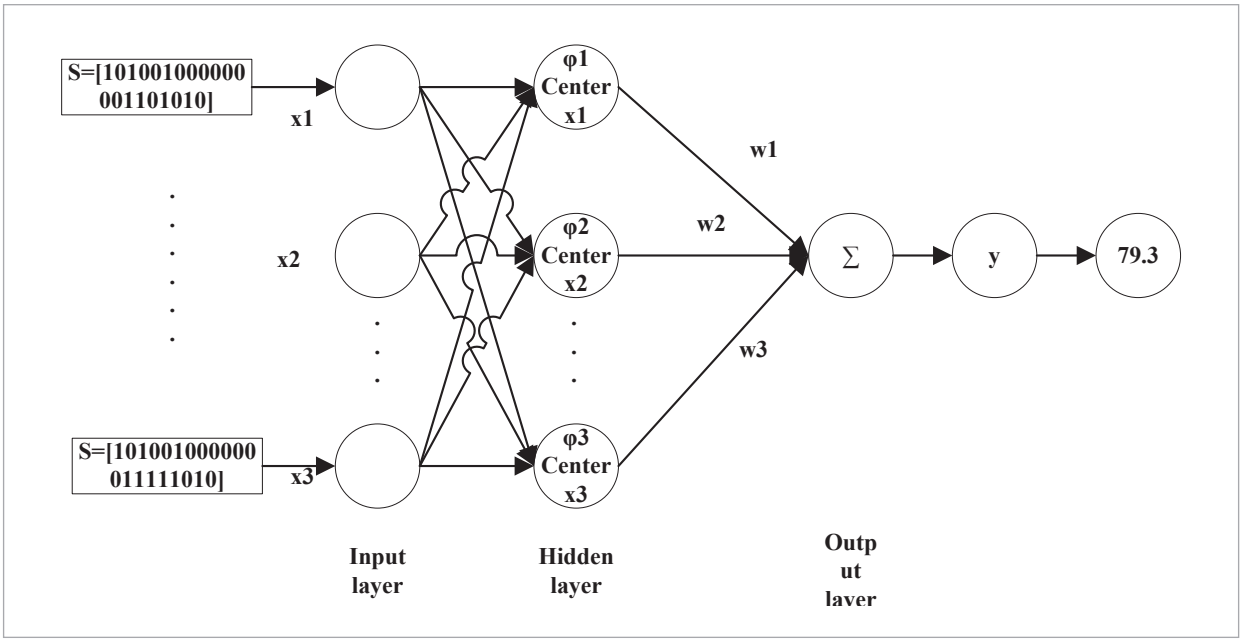
2.3 RBF Network Model

The weights of hidden layers are calculated by backpropagation, and linear or nonlinear functions [26] are used as the activation function of output neurons. In chord classification, the addition of the information valence dimension gives a greater probability of linear separation [17]. The model has some viability because there are a lot of classes in this study. It can utilize as many computation components in the hidden layer as there are training sample N. An RBF is used to represent each element, as displayed in Formula (15) below [18]. The core processing flow chart of RBF network model is displayed in Figure 1. The closer the distance, the larger the output value.

$$\varphi_j(x)=\varphi\left(||x-x_j||\right). \quad (15)$$

The vector x is the signal applied to the input layer.

Figure 1
Core processing flow chart of RBF network model



3 Comparative Experiment of Music Coordination Performance of Different NN Models

3.1 Experimental Environment

Based on the window 10 system, this experiment is implemented using the TensorFlow framework in Python. It uses an Intel Core i7-6800k Central Processing Unit with 16GB of RAM.

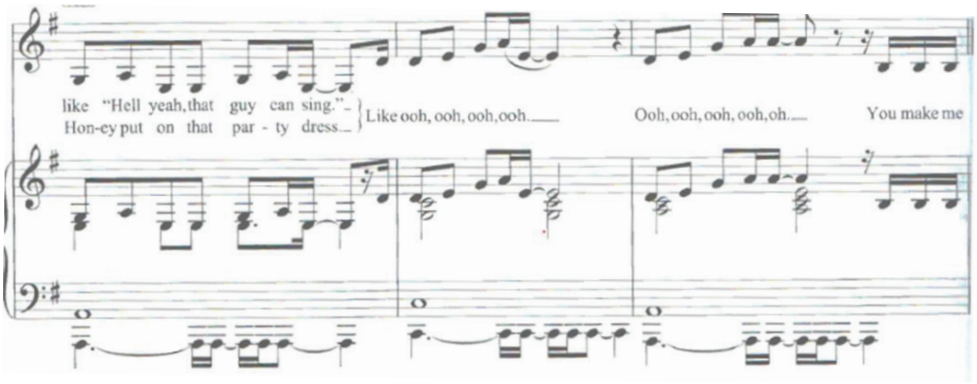
3.2 Experimental Data Set

This paper adopts three music datasets of Lim, Rhyu and Lee, including rock, pop, country, jazz, folk and other music genres. Each line represents a change in pitch or chord in music. A full song tonality, spiral root and chord kinds, note root, note length, etc., are all included in each song file along with time (the current beat of the song), measure (the measure in which the musical event is occurring), and other information. After standardized processing, 2154 songs were randomly selected in this paper, and the data set was divided into two parts using the ten-fold cross-validation method [19]. 80% is used as the training set, 20% as the test set, and some notes are displayed in Figure 2.

In order to deal with the diversity of songs in the database in terms of tone, rhythm and harmony, this paper standardized the music in the following ways [15]. Tonally, all songs were converted to c major in order to keep the message consistent. In terms of note timing and duration, the length of each note is multiplied by the reciprocal of the song beat to achieve the same total duration of music, while introducing emotional labels. They are respectively happy, sad, romantic, excited, lonely, neutral, and expressed as 001-110 in turn. The emotional label is represented at positions 6-9 from the back of the note vector, and the start time and duration of the other note are represented at positions 1-6 from the back of the note vector. In terms of chord types, in order to simplify the complexity of the system response, this paper focuses on major and minor triads, and outputs 24 chord categories. It produces a c_{hord} for each measure chord representing the note s.

$$n_{D\#}=[000100000000] \tag{16}$$

Figure 2
Shows part of the notes



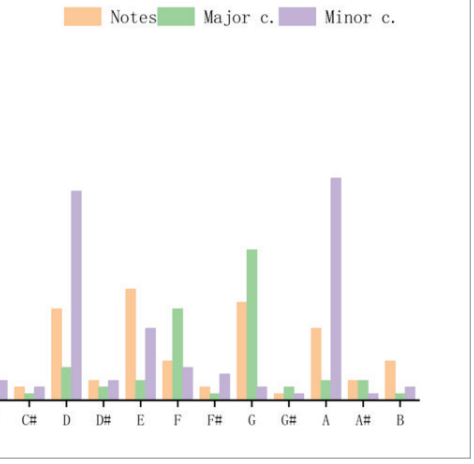
The note $D\#$ can be represented by the Formula (16), and the vector n_{ote} is used to represent a note. The chromatic scale is represented by 12 elements in this vector, along with a stop position.

Analogical chords are represented by the 24-position vector c_{hord}, 12 for major chords and 12 for minor chords. The two alternates with each other and the D major chord is displayed in Formula (17).

$$c_{D\,maj} = [000010000000000000000000] \tag{17}$$

There are 12 possible notes in the music, and only one chord is considered per measure. Test the percentage of notes and major and minor chords in the total set as displayed in Fig. 3. The 12 chords are C, C#, D, D#, E, F#, G, G#, A, A#, B. The vertical axis represents the proportion of chords to the total. The appearance of C, D, E, G and A is more obvious, and C and D occupy more than half of the total notes, respectively 65% and 51%.

Figure 3
The percentage of notes and major and minor chords in the test set



On the basis of the above, while considering that a bar adopts a chord and a variable number of notes and terminators, the design is to play three characters in the order of D, F and C. The encoding of each note would produce the following matrix's N_{NPMX12}, $N_{3\times12}$, which is displayed in Formula (18) below. NPM represents the note of each measure. At the same time, it simplifies and integrates the matrix, generates and vector S, as displayed in Formula (19) below. The bottom nine represent the emotional label happy, and the note starts at 5s and lasts for 2s.

$$N_{3\times12} = \begin{bmatrix} 001000000000 \\ 000001000000 \\ 100000000000 \end{bmatrix} \tag{18}$$

$$s = [101001000000001101010] \tag{19}$$

3.3 Experimental Process of Performance Comparison of Music Coordination

This paper uses five models to explore the performance of different NN models: LSTM RNN, LSTM RNN ensemble model of timing and SAM, LS-ESNs model, RBF network model, and multi-layer percep-

tron for performance comparison. Firstly, this paper standardizes the music sequence, and improves the fit of music melody and emotion by introducing emotion label. Then, in order to fully consider the order and duration of the note playing, the timing relationship of the note playing is introduced into the vector and. In the pre-training of the whole experiment, all samples in the training set were trained for 30 times (epoch=30), and the training times were dynamically expanded when parameters were adjusted. In this paper, a multiple of 10 is superimposed to 300 times (epoch=300). For the trained model, the total number of periods of 200 is set as the stop criterion, and a random gradient descent of $\eta = 0.001$ is used as the optimizer [28]. When the classification cross entropy reaches the minimum validation error, the optimal weight can be saved as a loss function. Finally, the five models were applied to the test set for verification, and the performance of the models was compared by comparing Accuracy, F1, k value, evaluation score and LOSS. The comparison of Loss in epochs 30 times and epochs 300 times is displayed in Figure 5 and Figure 6, and the experimental flow chart is displayed in Fig. 4.

In the experiment, the detailed settings of hyperparameters are shown in Table 1.

Table 1
Hyperparameter settings

Parameter	Value	Parameter	Value
Number of LSTM units	128	Learning rate	0.001
Number of LSTM layers	2	Batch size	64
Number of self-attention heads	4	Number of training iterations	1000
Number of hidden units in self-attention	64	-	

In this experiment, various parameters of the model were determined through repeated experiments and tuning. In the pre-training stage, all samples in the training set were first trained 30 times, and then the parameters were tuned by dynamically expanding the number of training times, and the number of training times was stacked to 300. The experiment used stochastic gradient descent as the optimizer, with a learning rate of 0.001, and the total number of epochs was set to 200 as the stopping standard.

In order to compare with other models fairly, the implementation is as follows.

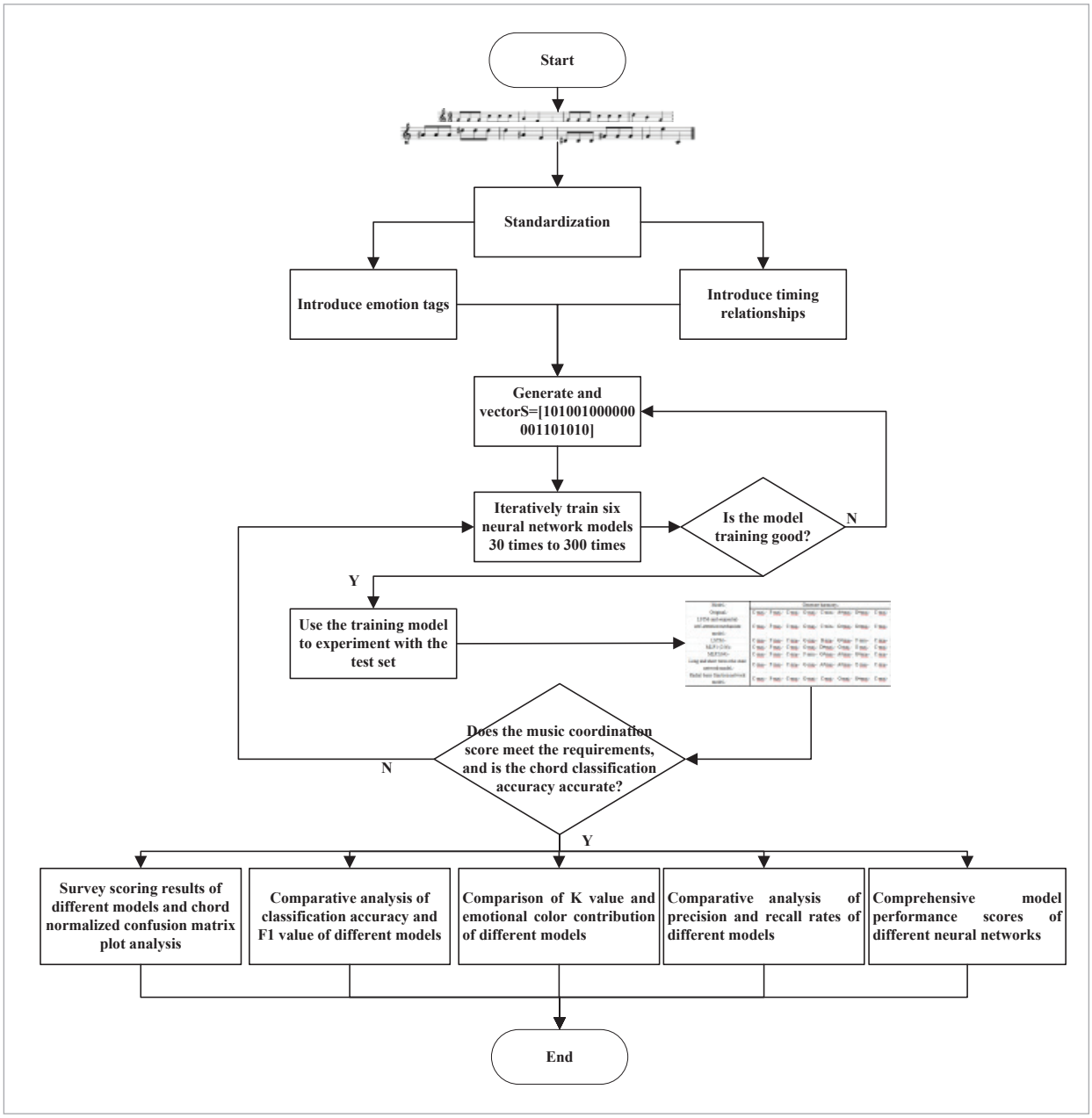
1. The experimental model is compared with other models in the same experimental environment, including operating system, hardware configuration and software framework.
2. All models are trained and tested on the same dataset to ensure the fairness of the comparison.
3. The same training strategy and parameter settings are used in the training process, including the number of training times, optimizer, stopping criteria and loss function.
4. The experiment uses the same evaluation indicators to compare different models, including accuracy, F1 value, K value, etc., and also comprehensively considers evaluation indicators such as loss function, expert evaluation score, and audience evaluation score to comprehensively evaluate the performance of the model.

In this study, the experiment includes a comparison of multiple models, among which LSTM performs well in processing time series data, can capture time correlation well, and is widely used in the field of music generation. The long short-term echo state network model is a recurrent neural network with memory ability. It effectively retains information when processing long sequence data and is suitable for tasks such as music generation. The radial basis function network model and multi-layer perceptron have certain advantages in processing nonlinear data and classification tasks, and may provide different perspectives on music coordination performance. The experiment adopts the comparison of these models to have a more comprehensive understanding of their performance in music coordination performance, and can provide reference and selection for different application scenarios. It can be seen that the comparison model is advanced and reasonable.

The criteria for comparison and selection of experimental models are as follows.

1. The comprehensive evaluation score is an indicator for evaluating the overall performance of the model, which comprehensively considers the per-

Figure 4
Experimental flow chart



formance of multiple aspects such as classification accuracy and music generation quality.

2 Music chord classification is an important aspect of ensemble music coordination, and classification accuracy is one of the important indicators for evaluating the coordination performance of the model.

3 The quality of generated music is one of the key indicators for evaluating the generation ability of the model. The study uses objective indicators to measure the quality of generated music and evaluates the authenticity, fluency and emotional expression ability of the model-generated music.

The reasons for selecting a specific architecture for comparison are as follows.

1 The architecture of the comparison model should have a certain similarity with the proposed model to ensure the fairness of the comparison. The model proposed in this experiment uses an LSTM network that combines time series and self-attention mechanisms. The comparison model should have a similar architecture to more accurately evaluate the performance difference between the two.

2 The comparison model has been proven to have certain advantages in previous studies or is applicable to similar tasks, which can improve the credibility and persuasiveness of the comparison results.

Figure 5 shows the loss function curves of different models after 30 iterations. The six NN models all converge to 1.3-1.5, and LSTM converges to 1.34 after 30 iterations. After 30 epochs of LSTM and sequential SAM model, the loss decreases from 2.54 to 1.31, which is 48% of the original decrease. Compared with LSTM, the loss is reduced by 0.03, LS-ESNs model reaches 1.35, and the effect is worse than that of LSTM and sequential SAM model, which requires 0.04 more loss. The other three models have worse effect overall. After 30 iterations, the loss of the six models has decreased after training and learning, and the models can learn music melody characteristics to a certain extent.

Figure 6 shows the loss function diagram of different models after 300 iterations. As a whole, the six models tend to 0.53 to 0.57, corresponding to the LSTM model, and the loss tends to 0.54 after 300 epochs. The MLP1(256) model has a minimum loss of 50.9%

Figure 5

Loss function diagram of different models after 30 iterations

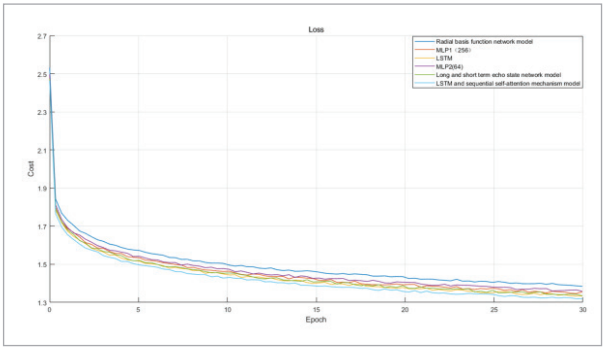
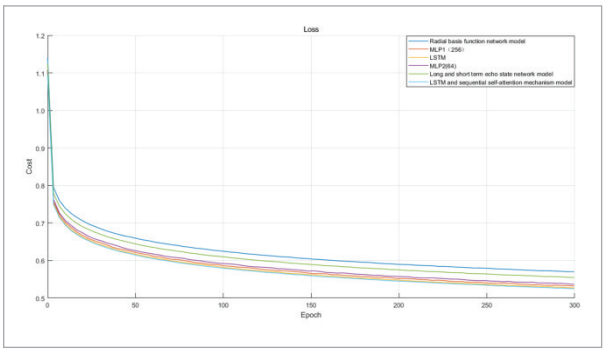


Figure 6

Loss function graph of different models after 300 iterations



from 1.12 to 0.55 over 30 epochals. After 300 iterations of LSTM and sequential SAM model, the model can fully learn the note features and timing relationships, and the loss is only 0.53. The minimum loss is 59.5% lower than the original 30 epoch approach and 0.01 lower than the LSTM model. The introduction of timing and SAM can reduce the loss to a certain extent, which is 0.02 less than MLP1(256) model, and the effect is better. RBF network mode has the highest loss, reaching 0.57, and the worst effect. By comprehensive comparison, the performance of LSTM model and LSTM and sequential SAM model is better, and it can achieve ideal results for the experiment.

4 Experimental Results and Discussion of Music Coordination Performance of Different NN Models

4.1 Experimental Results of Survey Scores of Different NN Models

A melody in America's music in the data set and its generated harmony are displayed in Table 2. LSTM and sequential SAM model produce the best harmonic effect.

In this paper, a total of 3000 questionnaires are used by randomly interviewing listeners and music experts. Real-time surveys determine whether the model can coordinate a piece of music, and score it. Table 3 depicted the typical results of the experimental survey scores.

Table 2
Comparison of generated and original harmony of different models

Model	Generate harmony							
Original	C maj	F maj	C maj	G maj	C min	A#maj	G#maj	C maj
LSTM and sequential SAM model	C maj	F maj	C maj	G maj	C min	G#maj	G#maj	C maj
LSTM	C maj	F maj	C maj	G maj	B maj	G#maj	F min	C maj
MLP1 (256)	C maj	F maj	C maj	G maj	D#maj	G maj	E maj	C maj
MLP2(64)	C maj	F maj	C maj	F min	G#maj	A#maj	G#maj	C maj
LS-ESNs model	C maj	F maj	C maj	G maj	A#maj	A#maj	E maj	C maj
RBF network model	C maj	F maj	C maj	G maj	C maj	G maj	G#maj	C maj

Table 3
Shows the typical results of some survey scores of different models

Assessment Score (100)	LSTM and sequential SAM model	LSTM	MLP1 (256)	MLP2(64)	LS-ESNs model	RBF network model
Expert 1	95	98	92	81	96	84
Expert 2	84	94	94	93	87	82
Expert 3	98	91	91	93	84	87
Expert 4	86	93	89	90	86	92
Expert 5	97	87	87	86	94	89
Audience 1	93	96	89	92	86	95
Audience 2	95	94	94	95	94	87
Audience 3	85	86	96	91	93	86
Audience 4	94	89	84	87	86	84
Audience 5	98	88	89	85	94	95
Total	925	916	905	893	900	881

4.2 Experimental Discussion of Music Coordination Performance of Different NN Models

4.2.1 Survey Score Results of Different Models and Confusion Matrix Graph of Chord Normalization

For different experts and different listeners, whether the model harmonizes the effect of a piece of music is judged as displayed in Table 2. Experts and listeners rated LSTM and sequential SAM model as the highest, reaching 925. The LSTM model is only second with a score of 916. The two models are more excellent in the processing of note melody, and can fully consider the temporal relationship and emotional tone between notes, so that listeners are satisfied. For the traditional multi-layer perceptron's MLP1 (256) is slightly worse, but has some feasibility in handling note melodies, while the RBF network model is the worst, with overall low ratings from experts and listeners. Overall, experts and listeners rated the model on how to coordinate a piece of music. This method is subjective to some extent, but it can provide some reference for the performance analysis of experimental models.

For different chords, the confusion matrix of chord normalization is displayed in Fig. 7. As can be seen from Figure 7, among the samples that actually belong to the corresponding actual category, G major

Figure 7
Confusion matrix diagram of chord normalization

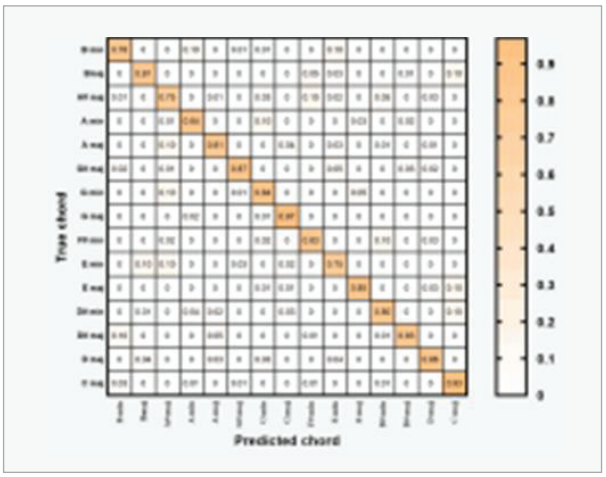
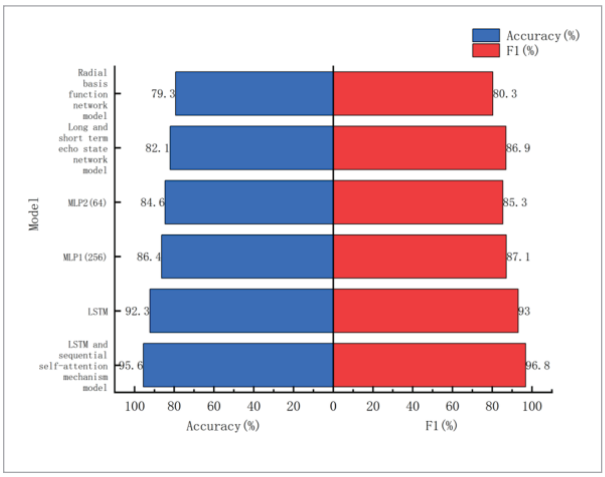


Figure 8
Classification accuracy and F1 values of different models



(G maj) has the highest correct prediction ratio, reaching 97%, showing a high classification effect. The classification effect of E minor (E min) and A flat major (A# maj) is poor, among which the correct prediction ratio of E minor is 75%. For E minor, 10% of the samples are incorrectly predicted as B major (B maj), 10% are incorrectly predicted as A flat major (A# maj), 3% are incorrectly predicted as G flat major (G# maj), and 2% are incorrectly predicted as G major (G maj). For other categories, the classification effect is good and can meet the experimental needs.

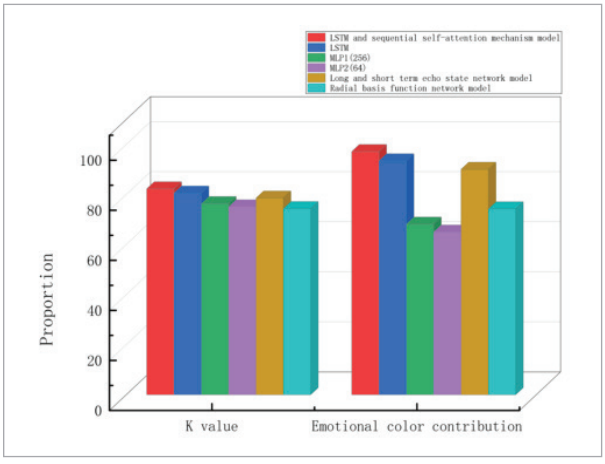
4.2.2 Classification Accuracy Rate and F1-Value of Different Models

Figure 8 shows the classification accuracy and F1 values of different models. In Figure 8, the highest accuracy is LSTM and sequential SAM model, reaching 95.6%. Compared with LSTM, it increased by 3.3%, and compared with LS-ESNs model, it increased by 13.5%, which is quite a big increase. The model has advantages in this experiment, and the accuracy rate for MLP1(256) is 86.4%, and the effect is slightly worse. Compared with MLP2(64), the RBF network model is better, with an improvement of 1.8%. The RBF network model has large errors, and cannot identify and classify chord types well. For F1 value, LSTM and sequential SAM model reached 96.8%, which increased by 3.8% compared with LSTM, with better effect, and increased by 9.7% compared with MLP1(256), with more accurate effect.

4.2.3 Comparison of K Value and Emotional Color Contribution of Different Models

For different models, the learning of emotional components is quite different, as displayed in Figure 9. Both LSTM and sequential SAM model and LSTM can learn emotional colors well, make full use of emotional labels, and coordinate moving music. The contribution of emotional color reached 97.2% and 93.5% respectively. In addition, for LS-ESNs model to a certain degree of coordination of pitch and musical emotional color, the performance was slightly worse, reaching 90.1%. At the same time, K value also affects

Figure 9
Comparison of K value and emotional color contribution of different models



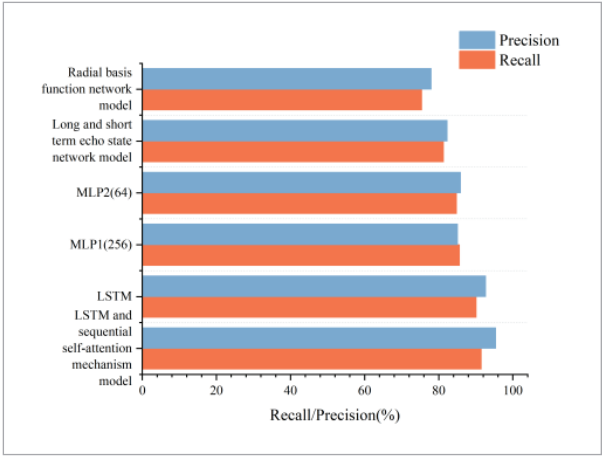
the degree of music coordination to a certain extent, and the K value of LSTM and sequential SAM model reaches the best effect, which is 82.3%. The K value of LSTM reached 80.6%, and for MLP1(256), the K value was only 76.4%. Compared with LSTM and sequential SAM model, the value decreased by 5.9%, and the effect was worse, while LS-ESNs model, to some extent, considered the timing relationship of notes, and the K value reached 78.6%, which was better than MLP1(256). In summary, the effect of LSTM and sequential SAM model and LSTM is relatively ideal, and it can play a good application in actual needs.

4.2.4 Precision Rate and Recall Rate of Different Models

The comparison of precision and recall rates of different models is displayed in Figure 10. For the recall rate, the LSTM reaches 90.1%, and the LSTM and sequential SAM model have the best performance, reaching 91.5%. Compared with LSTM, the effect of MLP1(256) is only second, reaching 85.6%. Its effectiveness is modest, but its complexity is low. The RBF network model has the worst effect, only 75.4%, and the effect cannot reach the ideal effect in the experiment. As for the precision rate, the precision rate of LSTM and sequential SAM model is the highest, reaching 95.4%, which is 2.7% higher than that of LSTM and 10.3% higher than that of MLP1(256). This shows the superiority of LSTM and sequential SAM model, with better performance.

Figure 10

Comparison of precision and recall rates of different models



4.2.5 Model Comprehensive Performance Scores of Different NN

In order to comprehensively compare the model performance of different NN, this paper uses Borda counting method to score and rank them, as displayed in Table 4. The total score of the experiment is 50 points, average to Loss, F1, K value, expert evaluation score, audience evaluation score. LSTM and sequential SAM model scores 43 points, accounting for a large proportion of the total score, reaching 86%, and LSTM scored 41 points. The total score of MLP1(256) is 36 points, which is seven points lower than that of LSTM and sequential SAM model, accounting for 72 percent of the total score. In addition, the overall effect of RBF network model is the worst, only 25 points, accounting for 50% of the total score, compared with LSTM and sequential SAM mode accounted for 36%. Therefore, under comprehensive evaluation, LSTM and sequential SAM model has the best effect.

Table 4

Model comprehensive performance scores of different NN

Experiment	Loss (10 points)	F1 (10 point)	K value (10 points)	Expert evaluation score (10 points)	Audience Assessment Score (10 points)	Total (50 points)
LSTM and sequential SAM model	9	9	8	9	8	43
LSTM	8	8	8	9	8	41
MLP1(256)	7	7	6	8	8	36
MLP2(64)	6	6	6	7	8	33
LS-ESNs model	5	5	5	8	7	30
RBF network model	5	5	4	5	6	25

5 Literature Discussion and Analysis

In this paper, we provide an in-depth understanding of the current research field of ensemble music coordination and the progress made by previous models in improving music classification accuracy, temporal coordination and emotion generation, but there are still some shortcomings in handling note order and duration and generating music emotion. A more critical analysis of the cited literature and a discussion of the potential for integration or improvement with the proposed model are now provided.

The IME model proposed by Clayton et al. is a method that integrates cultural shared knowledge and emotions to enhance the understanding and note coordination of IME in different music cultures [3]. Although this method has made some progress in cultural sharing, its modeling of note order and duration is still relatively simple and lacks in-depth consideration of music emotion. The model in this paper can draw on the cultural shared knowledge of the IME model and combine the timing and self-attention mechanism to better capture the long-distance dependencies and emotional changes in music. The RNN-based model proposed by Chakraborty et al. shows good results in human-machine collaborative performance, but has limited ability to generate emotions in music [1]. The model in this paper can inherit the temporal modeling ability of RNN and combine emotional labels with audio data to better express the emotional color of music and improve the quality of generated music.

Medina et al. used MLP to classify music emotions [14]. Although it achieved some success, the modeling of note order and duration was relatively weak. The model in this paper introduces timing and self-attention mechanisms to consider the order and duration of notes in more detail, thereby improving the quality of music generation and classification accuracy. Wood et al.'s research mainly focused on recording the movement data of musicians and did not directly involve music generation [20]. This study provides some valuable data references for this paper, which can be used to evaluate the coordination performance of model-generated music.

In summary, through the critical analysis of these previous studies, it can be seen that the integration or improvement potential of this model can be seen. By combining cultural shared knowledge, timing and self-attention mechanisms, as well as emotional labels and audio data, the model in this paper can comprehensively consider the cultural background, timing relationship and emotional characteristics of music, thereby improving the quality and coordination performance of music generation. This study is based on existing methods. By improving the model structure and feature integration method, it makes up for the shortcomings of existing methods and provides new ideas and methods for improving the coordination performance of ensemble music.

In this paper, a long short-term memory network (LSTM) and ensemble model based on the combination of timing and self-attention mechanisms are proposed to solve the problems of low classification accuracy, poor quality of generated music, and insufficient consideration of the order and duration of notes in music coordination. Compared with the traditional LSTM model, this model is more detailed in considering the order and duration of notes. At the same time, the introduction of emotional labels and audio data improves the quality of music generation.

There are many reasons behind the performance differences observed in the experiment.

1. The introduction of timing and self-attention mechanisms enables the model to better capture the long-distance dependencies and emotional changes in the note sequence, thereby improving the coordination performance of the model.
2. The method of combining emotional labels with audio data can make the model pay more attention to notes that match the target emotion, thereby improving the quality of music.

Among the components of the integrated model, timing and self-attention mechanisms are key components. The timing mechanism enables the model to capture the temporal relationship between notes in the time dimension, while the self-attention mechanism enhances the model's ability to automatically pay attention to the note sequences and features within the notes at different time steps. The combination of these two components enables the model to better understand the timing relationship and emotional changes in the music sequence, thereby improving the performance of the model.

The results of this study have important implications and potential applications in the field of music information retrieval.

1 By improving the music coordination performance and the quality of generated music, the model can provide a more accurate and expressive music representation for the music information retrieval system, thereby improving the quality and accuracy of the retrieval results. In the music recommendation system, the high-quality music generated by the model can be used to provide users with recommendations that are more in line with their personalized needs.
2 This study provides new ideas and methods for the field of music generation and creation. By introducing emotional tags and audio data, the model can generate more emotionally expressive and artistic music works, providing music creators with more diverse and personalized creative inspiration. The timing and self-attention mechanisms of the model can help the music generation system better capture the temporal relationship and emotional changes of music, thereby generating more coherent and rich music works.

6. Conclusions

LSTM network and ensemble model based on timing and SAM were used to study music. The paper used LSTM network to automatically learn important features of notes, introduced timing and SAM to enhance the model's ability to automatically pay attention to note sequences and features within notes at different time steps, and introduced emotional labels and timing data when standardizing music data. It combined emotional and temporal data information with audio data to make the model pay more attention to the notes and temporal relationships that match the target emotion, thus improving the music quality. This paper compared the performance of six kinds of ensemble models of LSTM and sequential SAM model, LSTM, MLP1 (256), MLP2(64), LS-ESNs model, RBF network model NN. It showed that LSTM and sequential SAM model had the highest accuracy of comprehensive evaluation score and chord classification, and the best effect. This experiment fully considered the order and duration of musical notes, synthetically compares six ensemble models of NN architecture, and explored the comprehensive performance of the models in music automatic coordination. This can improve the classification accuracy and coordination of the model, and has stronger practicability. However, there were some deficiencies in the experiment in this paper. The number of participants was a too large, and the data of the questionnaire was not enough. Subsequent experiments can be carried out to further optimize by enriching experimental data and lightweight methods.

The limitations of this study are reflected in the limitation of the dataset, the complexity of the model, and the subjectivity of the emotion label.

1 The music datasets used in the study include the Lim, Rhyu and Lee datasets. These datasets cover a variety of music genres, but there are still problems such as insufficient data volume and insufficient diversity of music styles, which affect the generalization ability of the model and make it perform poorly when processing a wider range of music genres.
2 The LSTM model based on the timing and self-attention mechanism has a relatively complex structure, a long training time, and high computing resource requirements, which limits its feasibility in practical applications, especially in resource-constrained environments.
3 The introduction of emotion labels improves the quality of music generation, but the definition and annotation of emotion labels are subjective. Different people may have different emotional understandings of the same piece of music, resulting in the music generated by the model in practical applications not fully meeting the user's emotional expectations.

Future research will be carried out from the following aspects: First, use a larger and more diverse music dataset to cover more music styles and types to improve the generalization ability and applicability of the model. Second, further optimize the model structure and training algorithm, reduce computing resource requirements, improve training efficiency, and study methods to simplify the model so that it can reduce complexity while maintaining performance and enhance the feasibility of practical applications. Third, study more objective and standardized emotion labeling methods to reduce the subjectivity of

emotion labels and enhance the emotional consistency and user satisfaction of model-generated music.

The experimental research results have great application potential in all aspects. In the music recommendation system, the model in this study can significantly improve the recommendation quality and user satisfaction of the music recommendation system. By generating high-quality music that meets the emotional needs of users, the recommendation system can provide more personalized and accurate recommendations. In automatic music generation and creation, the experimental model can be used in automatic music generation and creation tools to help music creators quickly generate music clips with high coordination and emotional expression, saving creation time and improving creation efficiency. In emotional computing and music therapy, the experimental model generates music that meets specific emotional needs and is applied in the fields of emotional computing and music therapy to help improve the user's emotional and psychological state and improve the treatment effect.

References

1. Chakraborty, S., Dutta, S., Timoney, J. The Cyborg Philharmonic: Synchronizing Interactive Musical Performances Between Humans and Machines. Humanities and Social Sciences Communications, 2021, 8(1), 1-9. https://doi.org/10.1057/s41599-021-00751-8
2. Chen, W. A Novel Long Short-Term Memory Network Model for Multimodal Music Emotion Analysis in Affective Computing. Journal of Applied Science and Engineering, 2022, 26(3), 367-376. https://doi.org/10.6180/jase.202303_26(3).0008
3. Clayton, M., Jakubowski, K., Eerola, T., Keller, P. E., Camorra, A., Volpe, G., Alborno, P. Interpersonal Entrainment in Music Performance: Theory, Method, and Model. Music Perception: An Interdisciplinary Journal, 2020, 38(2), 136-194. https://doi.org/10.1525/mp.2020.38.2.136
4. Daneshfar, F., Jamshidi, M. B. An Octonion-Based Nonlinear Echo State Network for Speech Emotion Recognition in Metaverse. Neural Networks, 2023, 163(2), 108-121. https://doi.org/10.1016/j.neunet.2023.03.026
5. Fan, J., Zhang, K., Huang, Y., Zhu, Y., Chen, B. Parallel Spatio-Temporal Attention-Based TCN for Multivariate Time Series Prediction. Neural Computing and Applications, 2023, 35(18), 13109-13118. https://doi.org/10.1007/s00521-021-05958-z
6. Fu, E., Zhang, Y., Yang, F., Wang, S. Temporal Self-Attention-Based Conv-LSTM Network for Multivariate Time Series Prediction. Neurocomputing, 2022, 501(1), 162-173. https://doi.org/10.1016/j.neucom.2022.06.014
7. Gauer, J., Nagathil, A., Eckel, K., Belomestny, D., Martin, R. A Versatile Deep-Neural-Network-Based Music Preprocessing and Remixing Scheme for Cochlear Implant Listeners. The Journal of the Acoustical Society of America, 2022, 151(5), 2975-2986. https://doi.org/10.1121/10.0010371
8. Gunawan, A. S., Iman, A. P., Suhartono, D. Automatic Music Generator Using Recurrent Neural Network. International Journal of Computational Intelligence Systems, 2020, 13(1), 645-654. https://doi.org/10.2991/ijcis.d.200519.001
9. Khare, S. K., Bajaj, V. Time-Frequency Representation and Convolutional Neural Network-Based Emotion Recognition. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(7), 2901-2909. https://doi.org/10.1109/TNNLS.2020.3008938
10. Leman, M. Co-Regulated Timing in Music Ensembles: A Bayesian Listener Perspective. Journal of New Music Research, 2021, 50(2), 121-132. https://doi.org/10.1080/09298215.2021.1907419
11. Li, J., Liu, X., Zhang, W., Zhang, M., Song, J., Sebe, N. Spatio-Temporal Attention Networks for Action Recognition and Detection. IEEE Transactions on Multimedia, 2020, 22(11), 2990-3001. https://doi.org/10.1109/TMM.2020.2965434
12. Li, Y., Liang, R., Wei, W., Wang, W., Zhou, J., Li, X. Temporal Pyramid Network with Spatial-Temporal Attention for Pedestrian Trajectory Prediction. IEEE Transactions on Network Science and Engineering, 2021, 9(3), 1006-1019. https://doi.org/10.1109/TNSE.2021.3065019
13. Li, Z., Tanaka, G. Multi-Reservoir Echo State Networks with Sequence Resampling for Nonlinear Time-Series Prediction. Neurocomputing, 2022, 467(1), 115-129. https://doi.org/10.1016/j.neucom.2021.08.122
14. Medina, Y. O., Beltran, J. R., Baldassare, S. Emotional Classification of Music Using Neural Networks with

the Mediaeval Dataset. Personal and Ubiquitous Computing, 2022, 26(4), 1237-1249. https://doi.org/10.1007/s00779-020-01393-4

15. Ray, J., Hendricks, K. S. Collective Efficacy Belief, Within-Group Agreement, and Performance Quality Among Instrumental Chamber Ensembles. Journal of Research in Music Education, 2019, 66(4), 449-464. https://doi.org/10.1177/0022429418805090
16. Sahoo, B. B., Jha, R., Singh, A., Kumar, D. Long Short-Term Memory (LSTM) Recurrent Neural Network for Low-Flow Hydrological Time Series Forecasting. Acta Geophysica, 2019, 67(5), 1471-1481. https://doi.org/10.1007/s11600-019-00330-1
17. Scalvenzi, R. R., Guido, R. C., Marranghello, N. Wavelet-Packets Associated with Support Vector Machine Are Effective for Monophone Sorting in Music Signals. International Journal of Semantic Computing, 2019, 13(3), 415-425. https://doi.org/10.1142/S1793351X19500028
18. Sharkawy, A. N. Principle of Neural Network and Its Main Types. Journal of Advances in Applied & Computational Mathematics, 2020, 7(1), 8-19. https://doi.org/10.15377/2409-5761.2020.07.2
19. Wong, T. T., Yeh, P. Y. Reliable Accuracy Estimates From K-Fold Cross Validation. IEEE Transactions on Knowledge and Data Engineering, 2019, 32(8), 1586-1594. https://doi.org/10.1109/TKDE.2019.2912815
20. Wood, E. A., Chang, A., Bosnyak, D., Klein, L., Baraku, E., Dotov, D., Trainor, L. J. Creating a Shared Musical Interpretation: Changes in Coordination Dynamics While Learning Unfamiliar Music Together. Annals of the New York Academy of Sciences, 2022, 1516(1), 106-113. https://doi.org/10.1111/nyas.14858
21. Xiao, Y., Yin, H., Zhang, Y., Qi, H., Zhang, Y., Liu, Z. A Dual-Stage Attention-Based Conv-LSTM Network for Spatio-Temporal Correlation and Multivariate Time Series Prediction. International Journal of Intelligent Systems, 2021, 36(5), 2036-2057. https://doi.org/10.1002/int.22370
22. Ye, Z., Chen, M. Visualizing Ensemble Predictions of Music Mood. IEEE Transactions on Visualization and Computer Graphics, 2022, 29(1), 864-874. https://doi.org/10.1109/TVCG.2022.3209379
23. Yeh, Y. C., Hsiao, W. Y., Fukayama, S., Kitahara, T., Genchel, B., Liu, H. M. Automatic Melody Harmonization with Triad Chords: A Comparative Study. Journal of New Music Research, 2021, 50(1), 37-51. https://doi.org/10.1080/09298215.2021.1873392
24. Yin, Z., Reuben, F., Stepney, S., Collins, T. Deep Learning's Shallow Gains: A Comparative Evaluation of Algorithms for Automatic Music Generation. Machine Learning, 2023, 112(5), 1785-1822. https://doi.org/10.1007/s10994-023-06309-w
25. Yu, Y., Si, X., Hu, C., Zhang, J. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. Neural Computation, 2019, 31(7), 1235-1270. https://doi.org/10.1162/neco_a_01199
26. Zhang, D., Zhang, N., Ye, N., Fang, J., Han, X. Hybrid Learning Algorithm of Radial Basis Function Networks for Reliability Analysis. IEEE Transactions on Reliability, 2020, 70(3), 887-900. https://doi.org/10.1109/TR.2020.3001232
27. Zheng, K., Qian, B., Li, S., Xiao, Y., Zhuang, W., Ma, Q. Long-Short Term Echo State Network for Time Series Prediction. IEEE Access, 2020, 8, 91961-91974. https://doi.org/10.1109/ACCESS.2020.2994773
28. Zheng, Q., Tian, X., Jiang, N., Yang, M. Layer-Wise Learning Based Stochastic Gradient Descent Method for the Optimization of Deep Convolutional Neural Network. Journal of Intelligent & Fuzzy Systems, 2019, 37(4), 5641-5654. https://doi.org/10.3233/JIFS-190861