Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Research on Pattern Generation and Structure Optimization of Neural Network Algorithms in AI Music Composition

Qi Liu[1,] ✉

[1] *Department of Art and Technology, Zhejiang School of Music and Dance, Communication University, Hangzhou, Zhejiang, 310018, China*

ABSTRACT

Artificial Intelligence AI composition is one of the hot topics that have been debated in recent years. In this paper, we first extract monophonic and chordal features from MIDI digital music files. Then the WaveNet intelligent music generation model is used as a carrier to optimize its multilayer convolutional network structure. The audio files are fed into the optimized WaveNet model, and the final training parameters are obtained after several rounds of iterative training. After the model completes the training, music sequences are automatically generated. The results show that the optimized WaveNet model for training leads to a significantly higher accuracy rate in the validation set than before optimization. Compared to other models, the method in this paper generates music using a larger variety of notes, improving the quality of the music theory and chord aspects. Compared with the composite scores of human compositions, the percentage of WaveNet model compositions with scores of 4 and above is about 20.3%, and the percentage of scores of 3 and above is 30.5%. Therefore, the overall level of the compositions generated by the model in this paper is good.

Keywords: waveNet model, MIDI digital music, convolutional networks, artificial intelligence

1. Introduction

In today's era of rapid technological development, the widespread application of artificial intelligence (AI) technology has brought about tremendous change and innovation in various fields. Music creation has always been an important field for artists to pursue expression and innovation. However, traditional music creation suffers from problems such as insufficient time and energy of creators, exhaustion of creativity, disharmonious musical connections, and simple and repetitive music, which limits the development and enhancement of music creation [1-3]. In the face of these problems, AI technology brings new possibilities for creators with the help of algorithms and big data analysis. On the one hand, AI technology can provide creative inspiration and creative suggestions by learning and

✉ Corresponding author.
E-mail address: 20140048@cuz.edu.cn (Q. Liu).
Received 15 January 2024; Revised 25 March 2024; Accepted 30 November 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-345
© 2025 The Author(s). Published by Combinatorial Press. This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

analyzing a large amount of music data, injecting new vitality into music creation, for example, AI technology can analyze the characteristics and laws of different types of music, helping music creators create more attractive and unique music works [4]. On the other hand, AI technology can realize the automatic generation of music scores and realize instant creation through generative modeling and automatic composition technology, which significantly improves the efficiency of music creation [5]. For example, AI technology can generate new musical works by learning a large number of musical works, enriching the content and form of music creation.

In fact, the origin of machine learning in music can be traced back to the end of the 20th century and the beginning of the 21st century [6]. During this period cross research in the fields of computer science, AI and musicology started to emerge with many ideas on how to use computer technology to understand, simulate and generate music [7]. Until the 2010s, deep learning, as a branch of machine learning, has seen significant development of its neural networks, which are mainly categorized into two main types: supervised neural systems and unsupervised neural systems for applications in music composition [8]. Neural network is a kind of training model that simulates the neuronal connection structure of the human brain and is widely used in music composition. Supervised neural networks can be trained to learn to generate music with specific musical features or styles, and the common ones include recurrent neural networks (RNN), convolutional neural networks (CNN), and long and short-term memory networks (LSTM), etc [9]. In music composition, unsupervised learning models can extract important information from unlabeled music data, reduce data complexity, and retain key music features, and excel in music clustering and classification, helping to discover different styles, genres, or patterns in music [10-11]. Unsupervised learning also supports transfer learning, i.e., the application of features learned from one musical style to the creation of another, commonly including self-encoders, generative adversarial networks, and variational self-encoders [12].

However, the AI music generated by the algorithmic technique in the previous research is only a simple style and melody imitation [13]. For this reason, some scholars have optimized this using neural networks. Zhang and Li [14] designed a music melody synthesis technique using RNN by extracting melodic features from existing audio in order to synthesize a new underlying melody, which can be synthesized with only static acoustic features. Whereas unsupervised neural networks are a class of neural networks that do not require labels or target outputs for training. Engel et al [15] added the NSynth note dataset to the WaveNet self-encoder model and implemented a process on the timbre produced by the instrument as a way of generating new sounds, but the difference between the span of notes and audio increased significantly. Jin et al [16] obtained a multi-track music generation model by learning the learning network through the Transformer model, obtaining sequences of piano, guitar, and drums, and fusing the three to obtain a multi-track music generation model, but its musical structure needs to be optimized. Kumar et al [17] used variational autocoder to optimize coherence and alignment in multi-track music generation and this model is superior to RNN model in synthesizing multi-track music. Shaikh and Jadhav [18] combined techniques such as Swish activation function and giant Trevally optimizer to design a dual-interactive Walthamian Fourier acquisition generative adversarial network, which optimizes the contradiction between lyrics and melody, emotion and structure. Nonetheless, AI music composition limitations remain, and how to use neural networks to optimize its pattern generation and structure under various compositional requirements is the key to improving the practical value of AI music.

The steps of AI music composition in this paper are feature extraction of MIDI files, instrumental feature model training and automatic composition note prediction. Firstly, the dataset is obtained by extracting the MIDI files for monophony and chords. The MIDI feature files were used as the initial input values for automatic composition, then the WaveNet model with multilayer convolutional network structure was selected as the training model, and the model structure was optimized by adjusting the five layers of convolutional layers and two layers of fully-connected layers to five layers of convolutional layers and three layers of fully-connected layers. Finally, the optimized WaveNet

model was used to automatically generate the specified note sequences. The effectiveness of the model in this paper is verified by analyzing and comparing the experimental data to measure the effect of composing.

2. Deep learning and neural network fundamentals

Deep learning is a branch of machine learning, machine learning is a branch of artificial intelligence, the concept of deep learning from the traditional neural network, but is not completely equivalent to the traditional neural network, but in the call, usually deep learning algorithms will contain the word "neural network", such as: recurrent neural network, convolutional neural network and so on. Deep learning is an upgrade of the traditional neural network foundation, the essence of which is roughly the use of human mathematical knowledge and computer algorithms to build an overall architecture, combined with a large amount of training data and large-scale computer computing power to adjust the internal parameters, and constantly adjusted to achieve the goal of the problem of the half-empirical, half-theoretical modeling approach.

2.1. Neural Network Basics

1) Neuron. A model, the network neuron model, was established in 1943 based on the study of biological neurons. The network neuron, or neuron, is the most basic unit that constitutes a neural network, and the mathematical function of a neuron from input to output is equation (1):

$$y = f\left(\sum_{i=1}^{n} W_i x_i + b\right) \tag{1}$$

W_i is the weight value corresponding to the input variable.

2) Activation function. The neural network is divided into three main layers: input layer, hidden layer and output layer.

3) Sigmoid activation function. The sigmoid function itself is a nonlinear smooth and differentiable function, which can be used in logistic regression to map the value of $(-\infty, +\infty)$ to the interval (0, 1), so it can fit the nonlinear problem.The expression of the sigmoid function is equation (2):

$$sigmoid(z) = \frac{1}{1+e^{-z}} \tag{2}$$

4) tanh activation function. The tanh activation function is another common activation function, and its principle is generally very similar to the sigmoid function, which utilizes an exponential function to perform a nonlinear transformation. Its definition and value domains are (-∞, +∞) and (-1, 1) respectively. Compared with the sigmoid function, the tanh function has some advantages, that is, its value domain is symmetric about the origin, but it also suffers from the problem of vanishing gradient, which also causes the training of the neural network to stop early.

The expression of tanh function is equation (3):

$$\tanh(z) = \frac{e^{2z} - e^{-z}}{e^{2z} + e^{-z}} \tag{3}$$

5) ReLU activation function. ReLU function is a segmented function, for the sigmoid function and tanh function in the existence of the gradient disappearance problem can be solved by it, its definition domain is $(-\infty, +\infty)$, the value domain is $(0, +\infty)$. ReLU activation function of the advantages of first of all is to be able to successfully solve the problem of the gradient disappearance, and secondly, it is very fast, only need to determine whether the independent variable is greater than 0 can be. However, it also has a problem that if the input value is negative, the eigenvalues will be eliminated directly. If the eigenvalues are eliminated, then the neurons with features less than 0 are no longer updated,

which will lead to neuron inactivation.The functional expression of the ReLU function is equation (4):

$$ReLU(x)=\begin{cases} z & if \quad z \geq 0 \\ 0 & if \quad z<0 \end{cases} \tag{4}$$

BP neural network, as the most commonly seen and used neural network we have nowadays, is likewise one of the most classical neural networks, and at the same time, it is also the basis of many other neural networks [19].

The training of BP neural network mainly has forward process and reverse process. In the forward process, the input vector is used as the input, which first passes through the hidden layer neurons, and the output of the j st hidden layer neuron is:

$$y_j=\alpha f\left(\sum_i w_{ji} input_i\right)+\beta \tag{5}$$

Where, α,β is the network parameters, w_{ji} is the weights of the i rd neuron of the input layer connected to the j th neuron of the hidden layer and f is the sigmoid function.

The output of the hidden layer is then used as the input to the neurons of the output layer and finally the output vector OUTPUT of the network is obtained with the formula:

$$output_s=\alpha g\left(\sum_j w_{sj} y_j\right)+\beta \tag{6}$$

Where, α,β is the network parameters, w_{sj} is the weights of the j rd neuron of the hidden layer connected to the s th neuron of the output layer and g is the sigmoid function.

In the inverse process, the gap between the target output lable and the actual output output is to be found first and represent it with the vector error, i.e.:

$$error=lable-output \tag{7}$$

The reason why the network needs to learn is because of the presence of the biaserror. The approach of the BP neural network is to compute the inverse of the error to each of the network's weights, and subsequently change the weights in the inverse direction of the inverse, which will reduce the change in the direction oferror. We use the error signal σ to represent the error information at the output of the network of the layer to be tuned. Backpropagation is the process of adjusting the weights of the network from the w_{ji} output in the direction of the input using error. First adjust the weights w_{sj} between the output layer and the hidden layer, and let the error signal of the s th neuron of the output layer be δ_s^0, there:

$$\delta_s^0=error_s g\left(\sum_j w_{sj} y_j\right) \tag{8}$$

And the error signal of the j st neuron of the hidden layer is δ_j^h, there:

$$\delta_j^h=f\left(\sum_i w_{ji} input_i\right)\sum_s \delta_s^0 w_{sj} \tag{9}$$

The layer to be adjusted is weighted according to the corresponding error, there:

$$w_{sj}^{(n+1)}=w_{sj}^{(n)}+\eta\delta_s^0 y_j \tag{10}$$

$$w_{ji}^{(n+1)}=w_{ji}^{(n)}+\eta\delta_j^h input_i \tag{11}$$

Where η is the learning step parameter.

Through this process, the neural network can more accurately achieve more accurate classification

and prediction by understanding and learning the features in the data.

2.2. Deep Convolutional Neural Network Structural Foundations

1) Fully connected layer. The fully connected layer is a network model with no connections within layers and full connections between layers, as well as a two-part graph model composed of neuron models.
2) Convolutional layer. Convolutional layer is the key structure of deep convolutional neural network. The essence of a convolutional layer is a set of filters which are parametrically trainable during the training process [20].

Each filter performs a convolution operation on the input data and outputs the convolved output through multiplication and accumulation operations. Since each image data is typically composed of three channels, the corresponding convolution kernel needs to be three-dimensional. The final result is obtained by simultaneously convolving and averaging each channel. The convolution operation results in a smaller size of the output result, and in order to ensure that the size of the image before and after the convolution remains the same during the actual convolution process, the padding technique is used to add zeros to the periphery of the input data in order to increase the size of the input data. In addition to this, the move step of the filter is stride, which determines the distance that the convolution window moves on the pixel matrix. The formula for calculating the size of the output features is shown below:

$$W_{out} = \frac{(W_{input} + 2 * p - F)}{s} + 1 \tag{12}$$

$$H_{out} = \frac{(H_{input} + 2 * p - F)}{s} + 1 \tag{13}$$

Where width and height are W and H, padding size is p, stride step is s and convolution kernel size is F.

For a single image input data, the number of channels of the output result of the convolution operation is the number of convolution kernels at the time of convolution. The pooling layer has fewer parameters compared to the fully connected layer, and can be stacked layer by layer, especially using the recently emerged residual pooling technique makes the depth of the model get a huge expansion, which largely improves the efficiency and effectiveness of feature extraction, and helps the model to be trained efficiently.
3) Pooling Layer. Pooling layer is the filter of convolutional feature extraction results, and its types are mainly divided into average pooling layer and maximum pooling layer. Pooling operation is the integration of the local characteristics of the features, which can be understood as a kind of dimensionality reduction, reducing the complexity of the data and attenuating the risk of overfitting.
4) Dropout. In order to optimize the model training, Hinton proposed the Dropout technique, which makes the neural network model in the training process, when each batch of data into the network, the model will be randomly deleted according to the set probability value of a certain proportion of neurons, so that the model uses less than the number of parameters than the original parameter for training, which improves the model's ability to generalize, and effectively avoids the overfitting, and then becomes One of the most commonly used techniques for neural network modeling.

3. Music pattern generation and structure optimization

3.1. Principles of Automatic Composition

For automatic composition, the creation of a piece of music is roughly divided into three steps: feature extraction of MIDI files, WaveNet model training, and automatic composition note prediction [21].

After feature extraction of the MIDI files, we obtained the monophonic and chordal sequences of all the instruments in the dataset, and the principle of automatic composition adopted in the project is similar to the model of prompting the upcoming words by the input method through the existing text. For this project, an array of monotone and chord sequences of length 32 is chosen as the input to the model, and the next P monotones or chords are predicted, and the prediction result is taken as the output, and then the output is taken as a part of the array of pitch sequences as the input to continue the prediction, and in this way a sequence of predictions is obtained, which is the resultant pitch data of the composition, and the principle of which is illustrated in Fig. 1. The principle is shown in Fig. 1.

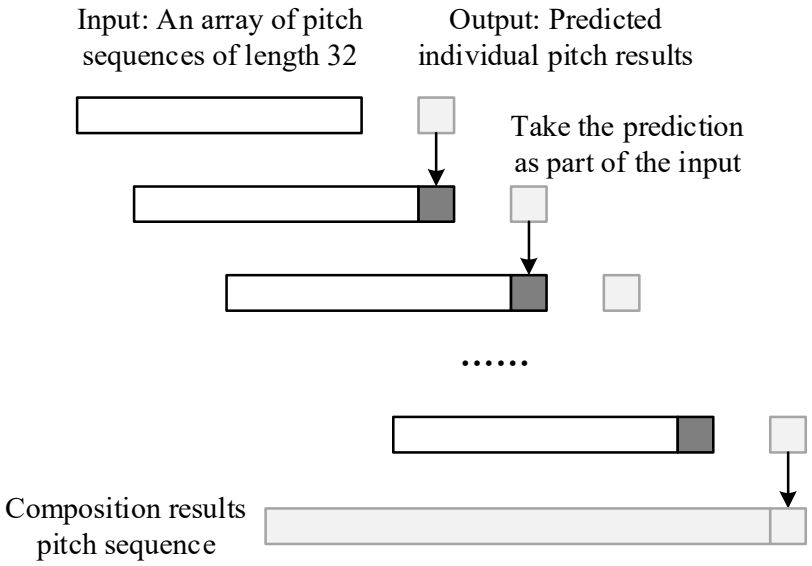


Fig. 1. Schematic diagram of automatic composition

3.2. MIDI feature extraction

The extraction of MIDI file features is essentially the extraction of individual musical tones and chords, because the tones and chords in the file are both the key inputs to our automated compositions and the resultant outputs of our musical compositions, and after that all the predicted pitches and chords are sequentially arranged and encapsulated into a track object, which results in the playing score of a single instrument [22].

3.3. WaveNet model

"Artificial intelligence AI composition" can be roughly regarded as an organic "imitation" of human composition behavior by artificial machinery. The biggest difference between this method of composition and the traditional composition process is that the entire process of artistic creation is completed by artificial intelligence independently, without the help of human external intervention. Users do not need to care about the process of "composing", and the focus of their operation is more on the process of "training", "simulation", and "algorithm". The emergence of artificial intelligence composition has brought a lot of impact to the music research community and academia. There are many music application platforms such as the above-mentioned ones, such as WaveNet developed by Google DeepMind, a system based on convolutional neural networks to generate music, WaveNet, which uniquely uses convolutional neural network systems to create, "WaveNet is different from some previous music creation applications, only for MIDI files or other simple music formats for parsing, WaveNet can directly target complete audio files (such as Wave, MP3, etc.), it has very strong expansion and audio processing capabilities, but the music generated by it lacks the inherent logic of music, and it is difficult to explain the logic with traditional music common sense [23].

The basic principle of WaveNet is to predict the value of the t nd point based on the first $t-1$ point of the sequence, so when we know the sequence of notes in a piece of music, we can use the model to predict the subsequent possible notes, the basic formula of WaveNet prediction is shown in equation (14):

$$p(x)=\prod_{t=1}^{T} p(x_t \mid x_1, x_2, ..., x_{t-1}) \tag{14}$$

The WaveNet model is a convolutional model with a multi-layer convolutional network structure, where each convolutional layer performs a convolutional operation on the data from the previous layer, and the larger the convolutional kernel is during the operation, the larger the perceptual range of the layer is, and the better the perceptual ability in the time domain is. When the output layer gets the final prediction result at the end of the convolutional layer operation, it outputs the prediction result and uses the result as part of the input for subsequent iterations in a continuous loop.

When the sampling rate of the audio is high, the convolutional layer is required to have a stronger perception of the time domain, so the Dilated convolutions model is proposed, which is optimized on the basis of the original WaveNet model, which adds the parameter dilation, and the value of dilation determines which nodes in the upper layer participate in the convolutional operation. Take the second convolutional layer as an example, at this time dilation is l, for the convolutional operation is only used to the tth, t-2, t-4, ... these points.

3.4. Structure of the automatic composition training model before optimization

In this experiment, the WaveNet model is used to compose music. The input of the model is an array of monophonic and chordal sequences of length 32, which is firstly convolved by multi-layer convolutional layers, and then predicted by a fully connected layer, and the predicted monophonic or chordal results are the outputs of the model, which are then compared with the next pitch or chord that actually occurs in the sequence of the dataset. Comparing the predicted result with the next pitch or chord that actually occurs in the data set, the cross entropy between multiple classifications is calculated, and the model parameters are adjusted by back propagation. In order to expand the perceptual domain of the model so that more pitches can participate in the prediction result, different expansion rates are set for each convolutional layer in the model.

Before optimization, the model consists of five convolutional layers and two fully connected layers, the convolutional layer takes the output of the previous layer as the input for the convolution operation, and after the operation of the convolutional layer 5, the result is sent to the fully connected layer, and after the operation of the fully connected layer, the prediction result is output, and then the cross entropy of each category is calculated to optimize the model parameters iteratively. However, in the training process, for some instruments with more data samples in the training set, the effect of training is not very good so consider optimizing the model.

3.5. Structure of the optimized automatic composition training model

In the research on the model and consult the relevant information to understand that there may be a variety of reasons that lead to the training set loss value delayed decline, combined with their own thinking, the main reason may be a single activation function, should increase the richness of the activation function pieces, in addition to this, the model of the number of fully-connected layer is too small, which may lead to a larger prediction error, can be appropriate to the fully-connected layer to be adjusted, and therefore the following optimization ideas are proposed:

1) Set the fully connected layers as 3 layers.

2) Set the number of neurons to 3000, 1024, and the number of single tones and chord types, respectively.

3) Set the activation function of the added fully connected layer to tanh activation function.

4) Set the previous 300 rounds of training to 150 rounds, with 30 as a checkpoint.

The model is adjusted accordingly, and after adjusting the model, the model structure is obtained as shown in Table 1.

Table 1. The automatic composition model structure after optimization

Input layer	Input layer	Enter the characteristics of each instrument
Hidden layer	Convolution layer1	The output channel is 64, the convolution kernel is 3, the convolution of the causal expansion is used, and the expansion of the convolution expands to 1, and the activation function is relu.
	Convolution layer2	The output channel is 64, the convolution kernel is 3, the convolution of the causal expansion is used, and the expansion of the convolution is 2active, and the activation function is relu.
	Convolution layer3	The output channel is 64, the convolution kernel is 3, the convolution of the causal expansion is used, and the expansion of the convolution is 4. the activation function is relu
	Convolution layer4	The output channel is 64, the convolution kernel is 3, the convolution of the causal expansion of the causal expansion, the expansion of the convolution is 8, the activation function is relu
	Convolution layer5	The output channel is 1, the convolution kernel is 1, the convolution of the causal expansion is used, the activation function is relu
	Full junction1	The number of neurons is 3000, the activation function is relu
	Full junction2	The number of neurons is 1024. activation function is tanh
	Full junction3	The number of neurons is single and chords, and the activation function is softmax
Output layer	Output layer	A single sound or chord of the output forecast

After getting the pitch array of a certain instrument, we first need to number the monotone and chords that appear in the instrument, and then use the number to replace the pitch information in the original array, and then divide by the total number of pitch and chord types so that all the numbering types fall in the [0,1) interval, and then the model training and prediction are carried out in turn, and the prediction results are output.

We know the model training process for automatic composing, and we have obtained a well-converged note prediction model, using which we can automatically generate a sequence of notes of our specified length using this process by giving an initial array of notes and letting him keep calculating them on his own, and then keep predicting the predicted results as part of the input.

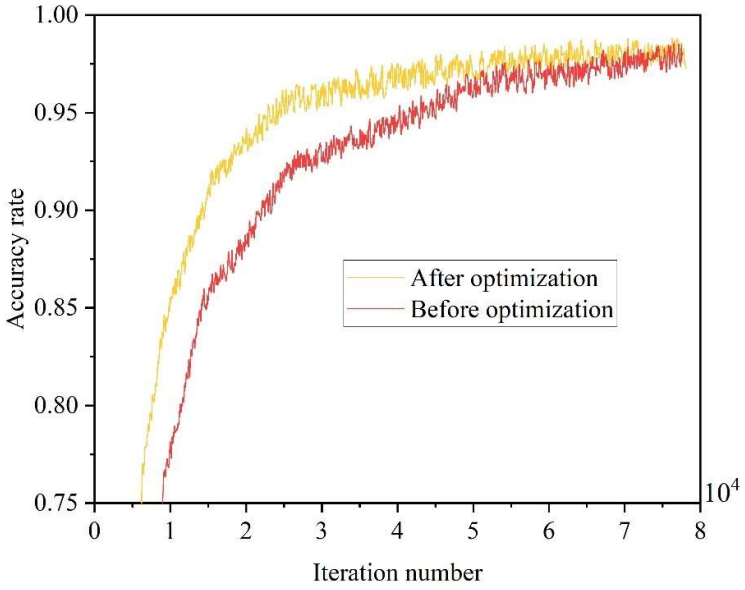
4. Music generation experiments and comparative analysis

4.1. Analysis and Comparison of Experimental Models

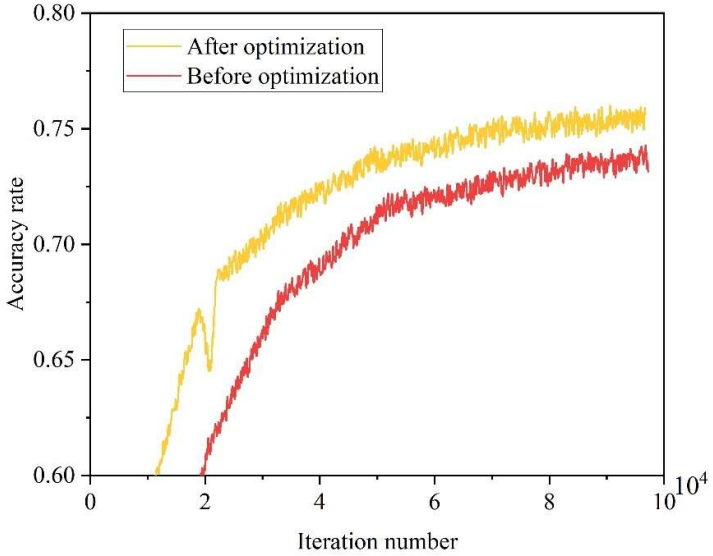
Accuracy rate is a commonly used evaluation metric to measure the percentage of output that is correct. It is difficult to say whether a particular note is correct or incorrect in an AI composition project, so this thesis defines the accuracy rate of AI composition as: the number of correct notes in the output / the number of all notes in the output, where a correct note means the same as a sample note in the test set. In order to ensure the fairness of the test, the accuracy rates of both datasets are calculated using the test set in the augmented dataset, which consists of a total of about 1,000 MIDI

piano compositions.

The accuracy curves of the training and validation sets are shown in Fig. 2. Although the fluctuation of the training using the pre-optimization WaveNet model is smaller, the function converges faster when using the post-optimization WaveNet model, and the accuracy is better with the same number of iterations. Also training with the post-optimization WaveNet model leads to significantly higher accuracy rates in the validation set than the pre-optimization one, proving its positive impact and feasibility for music creation.



(a)Accuracy of training set



(b)Verification set accuracy

Fig. 2. Accuracy ratio

5000 pieces of music were randomly selected from the ASMD database as training samples, and the same parameter settings were used to ensure the fairness of the experiment. Then, the Melody_LSTM network and the optimized WaveNet model were trained using the training samples, and after the training was completed, music generation was performed by letting each of the two models generate 300 pieces of music as test samples.

Twelve equal temperament is a method of music law, which divides an octave into twelve equal parts equally, and each equal part represents a semitone, and is the most dominant tuning method. This experiment mainly tests the use of notes in generating music by the two models; the more types of

notes used, the more diverse the music generated by the models. First, the number of times the twelve mean law notes are used in the two groups of test samples is extracted, and then, the percentage of the number of occurrences of each note is calculated to give a statistical distribution in the form of a proportion, and the results are shown in Figure 3. From the figure, it can be seen that the test samples generated by Melody_ LSTM have a high number of occurrences of C, D, E, G, A, and B tones, and the remaining notes are almost absent, and the distribution of notes in the test samples generated by the music composition model of the WaveNet model is relatively harmonic, i.e., the music composition model of this paper uses a greater variety of notes in generating the music, which indicates that the music composition model of this paper generates music with more varied tunes.

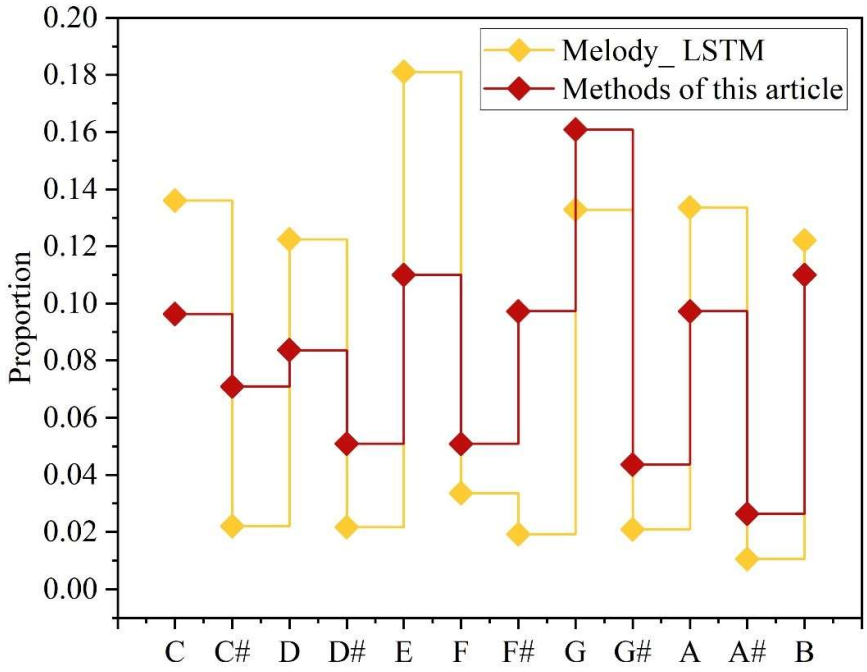


Fig. 3. Generate the music map

In order to test the effectiveness of the music generation model in this paper, the test samples generated by the WaveNet (pre-optimization) model are added to this experiment, and 300 pieces of music are generated as test samples using the same experimental environment configurations as the WaveNet (post-optimization) model. Seven effective feature information is extracted from the three sets of test samples, comparing the known music rules and calculating the statistical data, and the results are shown in Table 2. From the table, it can be seen that the music creation model in this paper effectively avoids the phenomena of excessive note repetition and interval span compared to the music generated by Melody_LSTM network, and the generated music has a very obvious improvement in classical style compared to the Melody_LSTM model. Meanwhile, the method in this paper (after optimization) is better than the pre-optimization in several indexes, which indicates that the use of optimized WaveNet model can improve the quality of the generated music in terms of music theory.

Table 2. Comparison of the rules of music theory

Feature	Melody_LSTM	Before optimize WaveNet model	After optimize WaveNet model
Over repetition	63.5%	40.1%	21.4%
Average self-correlation	0.15	0.09	0.04
The note is not in the tone	10.3%	6.8%	3.2%
The interval is less than eight	77.3%	81.4%	92.2%

degrees			
The biggest note of unique	60.4%	60.2%	65.7%
Unique smallest note	57.2%	61.8%	58.3%
Classical style	50.2%	62.2%	76.8%

In this paper, the concept of chord matching degree is proposed and experimented. The comparison results are shown in Fig. 4, the music chord matching degree generated by the music composition model (after optimization) model of this paper is overall higher than the other two, while the music chord matching degree generated by the music composition model (before optimization) model of this paper is not much different from that of the music generated by Melody_LSTM network.

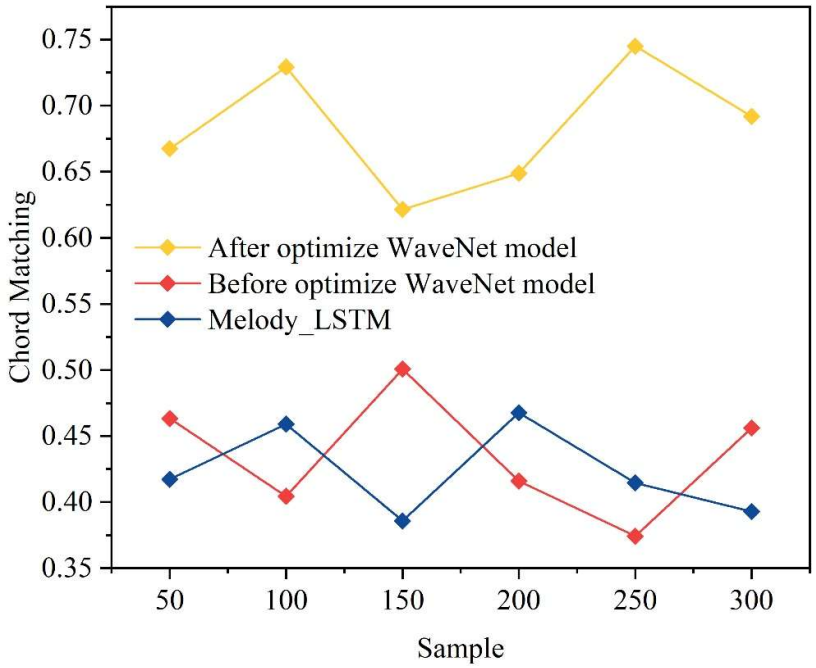


Fig. 4. Model chord matching

4.2. Composition results and evaluation

Music is a highly personal work, and in order to further validate the feasibility and effectiveness of the current algorithmic model, subjective evaluation was used and an online composition evaluation website was developed. The scores are selected from the scores created by composers in the training samples, the scores generated by the music composition model in this paper, and the scores generated by the basic_rnn algorithmic model in the Magenta project of Google Labs, and the scores are randomly disrupted in order to ensure the accuracy of the evaluation results. The auditor clicked on the left side of the score to audition, scored each index on the right side, and clicked submit to upload the scoring data to the server. Five music professionals, five ordinary people, and five music lovers were invited to participate in scoring 10 pieces of music created by the algorithmic model basic_rnn in the Magenta project of Google Labs, 10 pieces of music created by the design model of this paper, and 10 pieces of music created by composers.

In this paper, five evaluation indexes are designed: pleasantness, innovation, hierarchy, concordance, and stylistic clarity, and each index is set up with four grades, denoting {bad, average, better, and very good}, with corresponding scores of 1, 2, 3, and 4, respectively. The scoring results of a particular piece of music collected were recorded, and the scoring results are shown in Table 3.

Table 3. A list of music scores

Personnel	Melody	Degree of innovation	Hierarchical feeling	Concorde	Clarity of style
Major1	2	2	3	1	3
Major2	3	4	3	4	2
Major3	2	3	4	2	4
Major4	1	4	2	3	2
Major5	3	4	2	4	2
Ordinary1	2	3	3	2	3
Ordinary2	4	3	2	2	4
Ordinary3	2	4	3	4	2
Ordinary4	4	3	2	4	2
Ordinary5	2	3	4	2	4
Fondness1	1	3	4	2	2
Fondness2	2	3	4	2	3
Fondness3	3	3	4	4	2
Fondness4	3	2	4	2	4
Fondness5	2	2	1	3	3

The weights of each indicator were determined according to the subjective evaluation method, combined with the entropy weighting method. The process of calculating the weight of each index is as follows:

In the first step, all the indicators were standard normalized, and the standard normalized score of each indicator was obtained, as shown in Table 4.

Table 4. Normalized score

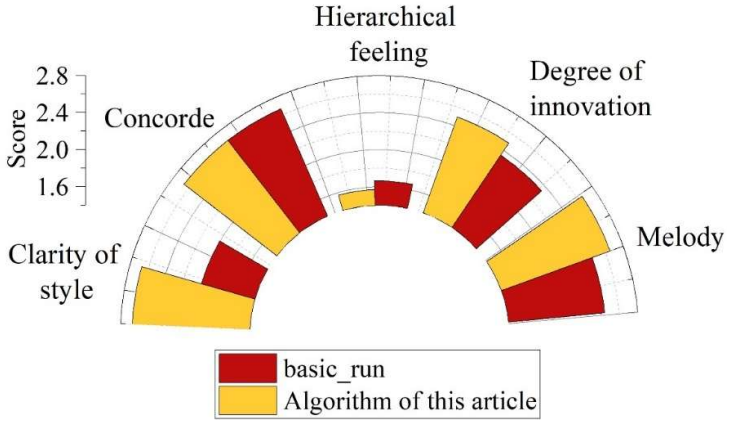
Personnel	Melody	Degree of innovation	Hierarchical feeling	Concorde	Clarity of style
Major1	0.33	0.33	0.67	0.00	0.67
Major2	0.67	1	0.67	1	0.33
Major3	0.33	0.67	1	0.33	1
Major4	0.00	1	0.33	0.67	0.33
Major5	0.67	1	0.33	1	0.33
Ordinary1	0.33	0.67	0.67	0.33	0.67
Ordinary2	1	3	0.33	0.33	1
Ordinary3	0.33	1	0.67	1	0.33
Ordinary4	1	0.67	0.33	1	0.33
Ordinary5	0.33	0.67	1	0.33	1
Fondness1	0.00	0.67	1	0.33	0.33
Fondness2	0.33	0.67	1	0.33	0.67
Fondness3	0.67	0.67	1	1	0.33
Fondness4	0.67	0.33	1	0.33	1
Fondness5	0.33	0.33	0.00	0.67	0.67

In the second step, the information entropy of each index is calculated, so as to get the weight of each index. The third step is to establish a fuzzy comprehensive evaluation matrix. Finally, multiply the fuzzy comprehensive evaluation matrix with the weights of each index to calculate the comprehensive score of a certain piece of music.

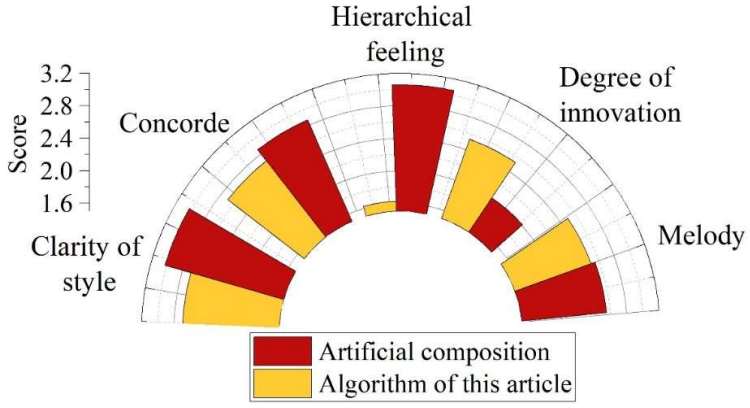
Among the 30 randomly selected pieces of music in this experiment, there are 10 folk songs written

by artificial composers, 10 composed by the design model in this paper, and 10 composed by the basic_rnn model in the Magenta project of Google Lab. Since the 5 parameter indicators are equally important, the scoring results of all the people on these music indicators are first averaged to get the average score of the 5 parameter indicators, and then compared separately. Indicator scoring results are shown in Fig. 5, which is the comparison results of 5 indicator scores of model compositions of this paper and model compositions of basic_rnn algorithm of Magenta project, which is the comparison results of 5 indicator scores of model compositions of this paper and people's work compositions.

The comparison results of the indicators show that the pleasantness and concordance of the compositions generated by this paper's model compositions and the basic_rnn algorithm model compositions are basically the same as those of the human work compositions, which indicates that the compositions can basically satisfy people's needs from the sensory point of view. From the point of view of style clarity, the music generated by the WaveNet model in this paper is higher than the music generated by the basic_rnn algorithm model. Generated music from the innovative degree, the network designed in this paper to learn the features of music from the training samples is neither too random nor dull. However, from the point of view of hierarchy, the music generated by the model in this paper has a large gap with the human work piece, the hierarchy of the music also affects the quality of the music, and the model designed in this paper still has room for further improvement.



(a) Basic_rnn and the algorithm



(b) Artificial composition and article algorithm

Fig. 5. Comparison result

In order to make an overall assessment of the quality of the generated music, all the compositions created by the human composers and the model of this paper were scored comprehensively, and the statistics of the comprehensive scoring results are shown in Fig. 6. From the final comprehensive

scoring results, 20.3% of the compositions generated by this paper's model reached 4 points, and about 30.5% of the compositions generated by this paper's model reached 3 points. This shows that the overall level of the music generated by the model in this paper is not bad, and can be comparable to human compositions, but the percentage of the lower value interval in the score is also higher than that of human compositions, which indicates that there is still room for further optimization of the composing network model designed in this paper.

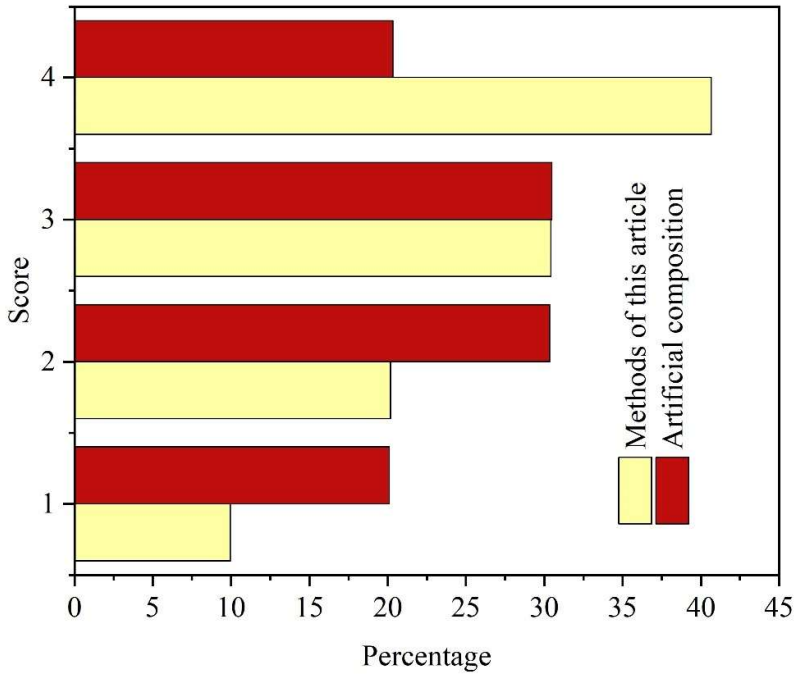


Fig. 6. Overall score

5. Conclusion

This paper focuses on AI automatic composition, using a convolutional model with a multilayer convolutional network structure, by training and optimizing the structure of its training model, and finally completing the composition of music. Conducting research for music intelligent generation, the main contents and contributions of this paper are as follows.

1) When performing training, the optimized WaveNet model converges faster and has a better accuracy rate compared to the pre-optimization one. When generating music, the music composition model in this paper has more note types and more varied tunes, and the use of WaveNet model can also improve the quality of the generated music in terms of music theory and chords.

2) 20.3% of the music generated by the model in this paper has a score of 4, which is a good overall level. Comparison of the scores of the five indexes of the compositions shows that the pleasantness and harmony of the compositions generated by the model of this paper are similar to the scores of the human compositions, and the clarity of the styles is higher than that of the compositions generated by the basic_rnn model. In terms of hierarchy, there is still a big gap between the scores generated by this paper's model and those generated by human compositions.

Funding

This research was supported by the Second Batch of 14th Five-Year Plan Undergraduate Teaching Reform Projects in Zhejiang Province in 2024: Exploration and Practice of Digital Teaching System for Music Design and Production Major Empowered by Artificial Intelligence.

References

[1] Huang, B. (2023). Modern music production equipment and how it can be used in student teaching: overture and its impact on motivation and interest in electronic music creation. Current Psychology, 42(34), 30499-30509.

[2] Matsunobu, K. (2018). Music making as place making: A case study of community music in Japan. Music Education Research, 20(4), 490-501.

[3] Ballico, C. (2017). Another typical day in this typical town: Place as inspiration for music creation and creative expression. Australian Geographer, 48(3), 349-363.

[4] Deruty, E., Grachten, M., Lattner, S., Nistal, J., & Aouameur, C. (2022). On the development and practice of ai technology for contemporary popular music production. Transactions of the International Society for Music Information Retrieval, 5(1).

[5] Gao, X., Chen, D. K., Gou, Z., Ma, L., Liu, R., Zhao, D., & Ham, J. (2024). AI-Driven Music Generation and Emotion Conversion. Affective and Pleasurable Design, 123(123).

[6] Sturm, B. L., Ben-Tal, O., Monaghan, Ú., Collins, N., Herremans, D., Chew, E., ... & Pachet, F. (2019). Machine learning research that matters for music creation: A case study. Journal of New Music Research, 48(1), 36-55.

[7] Albert, A. G. (2017). Computar Technology As A Means Of Enhancing Music Composition: Problems And Prospects. Business Journal for Entrepreneurs, (4).

[8] Cai, L., & Cai, Q. (2019). Music creation and emotional recognition using neural network analysis. Journal of Ambient Intelligence and Humanized Computing, 1-10.

[9] Cunningham, S., Ridley, H., Weinel, J., & Picking, R. (2021). Supervised machine learning for audio emotion recognition: Enhancing film sound design using audio features, regression models and artificial neural networks. Personal and Ubiquitous Computing, 25(4), 637-650.

[10] Schulze-Forster, K., Richard, G., Kelley, L., Doire, C. S., & Badeau, R. (2023). Unsupervised music source separation using differentiable parametric source models. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 1276-1289.

[11] Zhang, Z., & Chang, J. (2022). Clustering-based Categorization of Music Users through Unsupervised Learning. Economics & Management Information, 1-8.

[12] Chi, S., & Chen, H. (2025). Music style transfer and creation method based on transfer learning algorithm. Journal of Computational Methods in Sciences and Engineering, 14727978251318819.

[13] Verma, S. (2021). Artificial intelligence and music: History and the future perceptive. International Journal of Applied Research, 7(2), 272-275.

[14] Zhang, Y., & Li, Z. (2021). Automatic synthesis technology of music teaching melodies based on recurrent neural network. Scientific Programming, 2021(1), 1704995.

[15] Engel, J., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D., & Simonyan, K. (2017, July). Neural audio synthesis of musical notes with wavenet autoencoders. In International conference on machine learning (pp. 1068-1077). PMLR.

[16] Jin, C., Wang, T., Liu, S., Tie, Y., Li, J., Li, X., & Lui, S. (2020). A transformer-based model for multi-track music generation. International Journal of Multimedia Data Engineering and Management (IJMDEM), 11(3), 36-54.

[17] Kumar, N., Surendran, R., & Sumathy, K. (2024, November). Improved AI-Generated Multitrack Music with Variational Autoencoders (VAE) for Harmonious Balance Compared to Recurrent Neural Network for

Coherence. In 2024 4th International Conference on Advancement in Electronics & Communication Engineering (AECE) (pp. 1281-1286). IEEE.

[18] Shaikh, T., & Jadhav, A. (2025). Music generation using dual interactive Wasserstein Fourier acquisitive generative adversarial network. International Journal of Computational Intelligence and Applications, 24(01), 2450026.

[19] Lingbin Shen,Liping Tian,Hongbing Yao,Dongpeng Tian,Yifan Ge,Zhongmou Sun & Yuzhu Liu. (2024) .Rapid detection and recognition of phosphors using laser-induced breakdown spectroscopy and principal component analysis method—back propagation neural network algorithm. Laser Physics,34(5).

[20] J.D. Camacho,Carlos Villaseñor,Javier Gomez Avila,Carlos Lopez Franco & Nancy Arana Daniel. (2025) .Auto-compression transfer learning methodology for deep convolutional neural networks. Neurocomputing,630,129661-129661.

[21] J.G. Michopoulos,A.P. Iliopoulos,C. Farhat,P. Avery,G. Daeninck,J.C. Steuben & N.A. Apetre. (2024) .Bottom-up hierarchical and categorical metacomputing for automating composition and deployment of directly computable multiphysics models. Journal of Computational Science,79,102295-.

[22] Oscar Gomez Morales,Hernan Perez Nastar,Andrés Marino Álvarez Meza,Héctor Torres Cardona & Germán Castellanos Dominguez. (2025) .EEG-Based Music Emotion Prediction Using Supervised Feature Extraction for MIDI Generation.Sensors,25(5),1471-1471.

[23] Supriya Sridharan,Swaminathan Venkataraman & S. P. Raja. (2025) .SOC Estimation in Lithium Ion Batteries using a Hybrid Gated Residual Wavenet with LSTM-GRU Network Across Varying Temperatures. Journal of The Electrochemical Society,172(3),030510-030510.