An Improved Music Composing Technique Based on Neural Network Model

Mingheng Liang

CITI University, Ulaanbaatar 999097, Mongolia

Correspondence should be addressed to Mingheng Liang; 41823038@xs.ustb.edu.cn

Received 7 April 2022; Revised 23 May 2022; Accepted 10 June 2022; Published 11 July 2022

1. Introduction

Music creation refers to the complex mental and technical production process of music professionals or composers to create music with musical beauty. The main method is to combine different syllables according to the corresponding time series, such as melody and harmony, to produce dynamic sound waves with special timbre and texture. Music creation is usually created by composers who have received professional music training and education. It is an extremely complicated technique and task [1]. Music, on the other hand, is also a complicated artistic profession that may purify the hearts of people, illuminate intelligence, and balance feelings, and it has progressively become an essential element of our life [2]. To address the growing need for musical labour, a growing group of experts and academics are investigating music-generating modes. The expressive styles of music have become increasingly diverse as society and science and technology forces advance, and the integration of computers and music encourages the rapid growth of the music production area. In the past, people’s working music required important professional music theoretical knowledge in the fields of harmony, composition, rhythm, and beat. Since the 20th century, computers have played an essential part in all elements of music creation, and algorithmic composition has been kept as a technical term to characterize the process of composition and creation that uses algorithmic methods as the primary means of formation [3].

Digital art is a type of art that is generated and saved digitally on a computer, usually in binary form. The method is a massive undertaking in which the computer is merely a knowledge storage medium or a tool for generating works of art. The classification of digital media art is not technically differentiated according to the standards of digital media art. It is classified based on the primary domains or related categories of creative endeavors in which digital media technology is used [4]. The technical techniques of composition examined in the study [5] are used to represent the composer’s musical concepts through the use of fundamental musical theory, acoustics, polyphony, fitting method, and musical form structure. The practicability of system design, as well as the concept and evolution of computer music, must be examined in the discussion of algorithmic
composition. From a macro point of view, the algorithmic composition is a process in which creators use algorithms to analyze and program music fragments, elements, or the underlying laws contained in them, and drive the computer to generate music works [6].

With the emergence of machine learning, several scientists have attempted to use AI technology in music composition and have achieved substantial progress. Machine learning composing significantly reduces the threshold for music production, allowing nonprofessionals to contribute despite they do not know how to play musical tools. Meanwhile, it enhances the effectiveness of expert song composition and gives music composers a wealth of music materials. Machine learning composition has a variety of applications and may be seen in marketing, social interaction, entertainment, and other situations [7]. Deep learning techniques with improved artificial intelligence performance in image identification, video detection, natural language processing, and audio processing are being explored rapidly [8]. Furthermore, since it has a wide variety of applications in associated sectors, the research and enhancement of deep learning model technology and application scenarios are becoming increasingly significant. The deep learning algorithm is a new dimension reduction algorithm for the multilayer neural network. For the scene tasks of music creation and generation, deep learning can effectively construct music datasets and select models to generate new music [9]. This will also make music creation more accessible to more people and bring more beautiful music of different types and styles to human beings. For example, the earlier study of [10] builds a generation model based on neural network results, which can generate harmonious and beautiful music like that created by human beings. The model based on the structure design of the RNN deep neural network in literature [11] generates Pop Music by combining music with prior knowledge. In the work of [12], sequence modeling is carried out through a neural network, and auxiliary music creation is carried out based on simple music data sample set. Besides, in the work of [13], chord music is generated through bidirectional short and long-term memory (BLSTM) network structure model and corresponding dataset. Additionally, in the work of [14], dance music generation is carried out by a weakly supervised deep recursive neural network with audio energy power spectrum as input. Similarly, in [15], based on the self-coding machine (VAEs) and adversity-generation network model (GANs) in deep learning, the generation of music lattice migration is carried out.

At present, artificial intelligence is still in the era of weak artificial intelligence, how to effectively input human music knowledge and system into a deep neural network for artificial intelligence learning has become the core problem in this field of research. As a result, one of the current research priorities is to figure out how to efficiently preprocess massive amounts of music data to provide the network with relevant data. This study deals with the topic of algorithmic composition and the extension of artificial intelligence deep learning to classical algorithmic composition, based on the deep learning thinking mode and the preprocessing of music data based on music theory. This study will be targeted to study music creation of three main elements: melody, duration, chords, its data processing in the system in this research, and how to use the depth of the neural network learning method, neural network discovery data set of deep abstract rule, and use these weights independently produce characteristic of the music.

This article presents the first effort to apply a neural network model for music composing. For efficient composing, this article firstly segments the music set into a music sequence based on unit duration and extracts the Meier cepstrum coefficients of music audio during preprocessing. Secondly, it processes the training samples, which are made up of feature vectors from data that has been learned and predicted using short- and long-term memory models. Finally, the generated music sequence is spliced and fused to get new music. According to the experimental results, this model has the highest accuracy of 99% and the lowest loss rate of 0.03.

The rest of this article is organized in such a manner that Section 2 is based on related work, Section 3 explains our proposed methodology and model, experimental work and analysis are presented in Section 4, and conclusions and future of this work are presented in Section 5.

2. Related Work

Currently, music composition has captured the interest of members of the scientific community and has received a great deal of attention. Musical composting is the act of conceptualizing a piece of music or the craft of generating music and the final output is referred to as musical composition. A single note has no meaning. A song may be split into multiple portions in terms of music theory, and each section is made up of a succession of notes. As a result, a musical section is the most fundamental unit for conveying meaning. Music emotion and meaning value can only be represented by organically arranging these components. In music, time is split into equal parts, each of which is referred to as a beat. The beat is used to record the length, pace, and intensity. For example, the link between the downbeat portions is 3/4 of the melody. There is a beat with a quarter note in each sector, and each section includes three shots, strong or weak, in a series of strong and weak, and the weak. When the piece’s specified pace is 60 beats per minute, the length of each beat is 1 s and 3 s is a measure. This kind of unit music is a meaningful and melodic music mode, so the problem of music creation becomes the problem of reorganization of unit music.

The traditional research used a support vector machine to classify audio scenes through training, and the results proved that the classification accuracy was higher when MFCC was used as a feature vector. In [16], the authors developed an audio feature extraction system YAAFE under GNU (General Public License) for fast audio feature extraction. Furthermore, AI composition is to score as the carrier, the essence of research for the text mining research, first proposed in this article to the audio itself as the research object, from the perspective of the MFCC, audio signal
processing, and AI composer fusion, is proposed based on LSTM-RNN music audio automatic synthesis algorithm. AI composition is verified with the possibility of audio as the carrier, making the results more visible to the audience. The authors of [17] developed a tripartite tonal representation of the verse and chorus. On this basis, the earlier work of [18] also used recurrent NT neural network (RNN) to learn classical music. Their research first reorganizes the neural network learning music clips and then compares the original Bach music compositions to the usage of a neural network utilizing fragments of music to compose music. In terms of validation, the multiple classification test indexes test the result of the experiment, the final test results show that it has a larger gap with the human senses.

On other hand, composition based on music grammar requires the composer to model the rules of specific domain knowledge and compose according to the rules of modeling. Music, like human language, has its syntactic structure. Notes in music can be regarded as words, and grammar in the composition is defined by the composer himself [19]. The disadvantage of this method is that it is affected by the background of rule and creator knowledge modeling. The rule-based knowledge system is based on the rules of the music knowledge system. By establishing a rule set, the computer generates a melody based on the rule set after generating music, so that the generated music is more in line with the music rules. But because the rules of music are so complex, it will be a huge task to distill them all. An artificial neural network is a mathematical model that imitates the human brain and has a strong learning ability. Compared with other algorithm models, the creation of an artificial neural network mainly finds patterns and rules of composition by analyzing and learning a large number of materials and then creates based on this pattern. Although there is still a big difference between the machine’s creative ability and aesthetic ability of art and human beings, with the continuous improvement of science and technology, the neural network will have a greater space for development in the field of music. It is not just in recent years the neural network composition has emerged, in 1994, Michael C. Mozer et al. designed concert, a composition software system, using recurrent neural network (RNN). The system takes the recurrent neural network as the composing network structure. By inputting a large number of works for learning and training, the characteristics of music style are extracted and new music works are generated on this basis. Eck and Schmidhuber summarized the shortcomings of previous use of recurrent neural networks for the creation and proposed the use of an improved long short-term Memory (LSTM) network of recurrent neural networks to generate melody with a specific style [20]. Compared with the previous cyclic neural network, LSTM can solve the problem that the structure of the cyclic neural network cannot remember the long-distance information. Inspired by the preceding, this article developed a music composition approach that used a neural network and achieved 99–100% accuracy throughout the music composting process.

3. Methodology

This section explains the methods of the proposed study, such as audio prediction, music synthesis, our model training, and its prediction. Our proposed model architecture is made up of three elements. An input module is the first module. The query module is the second module, while the retrieval module is the third. Each module focuses on completing certain activities to gather and compose data. The suggested system’s concept is built on the interdependence of the input and query modules [21]. The person who submits the inquiries is also the one who records the composing keywords that regulate the retrieval process. The design of our music composition system is depicted in Figure 1.

The input module’s major purpose is to prepare the groups of music composing data that will be obtained from the query module. Every collection is focused on a single topic and is classified using a set of audio keywords. All audio input keywords are improved using particular algorithms to eliminate silence and noise. The query module’s primary responsibility is to formulate and refine queries given by the user. The techniques used to improve the question in the Query module are identical to the strategies used to improve training data in the input module. The MFCC characteristic is retrieved from the supplied query to provide testing data, which is then compared against training data in the retrieval module. The retrieving module’s major purpose is to compute the similarity among training and testing data to recover music composition recordings.

3.1. Audio Prediction and Music Synthesis. The proposed model is based on a recurrent neural network (RNN), which is developed from an artificial neural network. Due to its special structural characteristics, it has a short-term memory function for time series. The cycling unit of the neural network with the feedback is the place where feed-forward neural network is quite different, in addition to receiving input from the outside world. Furthermore, it receives the moment before the hidden state information, which may be viewed as a memory for a moment before, whereas feed-forward neural networks, due to structural limitations, can only accept input from the outside world. Therefore, it cannot be used to process speech and other time sequence signals closely related to the signal before and after. Recurrent neural networks have been widely used in video, speech, text, and other time sequence-related problems.

3.1.1. Split Unit Music. In collecting musical units, the goal is to maintain the strength of the music’s beat and the brief melody. As a result, if the value of the unit of time \( t \) is too little, the section’s integrity might be destroyed, and the strength of the music rhythm sensation is lost. However, if the unit of time \( t \) value is too large, it is easy to store too much melodic information, hence this article uses the unit of time \( t = 3 \) s in the test. When the music tempo is between 90 and 180 beats per minute, the number of bars per music \( m \) is
between 2 and 3 bars. Furthermore, in audio coding, coding stream $d_m$ is dependent on duration. According to music duration, the audio stream is cut into an audio fragment sequence of equal length. Equation (1) is used to cut stream data $d(t)$.

$$d(t) = d_m[0: f_{s}\ast t],$$  \hspace{1cm} (1)

where $t$ is unit time and $f_{s}$ is the sampling frequency of the audio file.

3.1.2. Feature Processing. The MFCC feature processing approach consists of windowing the signal, performing the DFT, obtaining the log of the amplitude, and then bending the frequencies on a Mel scale, followed by using the inverse DCT. Music conveys emotional information by affecting people’s auditory perception. Experiments show that the auditory perception has a linear change in the change of pitch. In addition, MFCC reflects the pitch hearing characteristics of the human ear through the logarithmic transformation of frequency and tone. The research results of music emotion and scene classification with audio as the carrier show that MFCC can efficiently identify the tone and frequency of music signals, which can be used as features of audio classification [22]. Therefore, this article takes MFCC as the feature of unit music [23].

The common MFCC has 39 dimensions, which are composed of 13-dimensional static coefficient. Among them, the difference coefficient represents the dynamic characteristics of music, while the 13-dimensional static coefficient is composed of one-dimensional energy characteristics and 12 sustaining numbers. The calculation process of MFCC can be explained in Figure 2.

Step 1. Perform Fast Fourier Transform (FFT) to calculate the amplitude spectrum of each frame signal.

Step 2. Transform the amplitude spectrum to Mayer domain using Mayer scale, and after filtering by mayer filter banks of equal bandwidth, superimpose the output energy of the filter banks:

$$S_k = \log\left(\sum_{j=0}^{N-1} H_k(j)|X(j)|\right), \hspace{0.5cm} k = 1, 2, \ldots, K, \hspace{1cm} (2)$$
3.2. Model Training and Prediction. The training sample is expressed as \((V\text{(pre}(m_i)), V(m_i))\), let \(S = \{M_1, M_2, \ldots, M_n\}\) of the dataset containing \(n\) pieces of music \(m\) and \(I\) is the index of the unit audio \(M_i\) in the dataset \(S\). Therefore, for this model, the input is \(\text{pre}(m_i)\), the pre-sequence music sequence of unit audio \(M_i\), in the form of \([V(m_1), V(m_2), \ldots, V(m_{i-1})]\), and the output is the similar feature vector \(H\) of unit audio \(M_i\). Here, \(M_i\) is determined by calculating the distance between \(H\) and unit audio in data set \(S\). The objective function of this model is set as \(\text{tanh}\) function, LSTM-RNN model music prediction problem \(F(\text{pre}(m_i); \theta)\). The problem can be expressed as a function construction problem of parameter set \(\text{alpha} = (W, U)\) and can be calculated by

\[
F(\text{pre}(m_i); W, U) = h_i, \quad (5)
\]

\[
h_i = o_i \cdot \text{tanh}(c_i), \quad (6)
\]

where \(o_i\) represents the output gate in the LSTM model and \(V_i\) represents the music vector \(V\) (pre\((m_i)\)) of \(\text{pre}(m_i)\) at the \(i\)-th moment. The output gate in the LSTM model can be given as

\[
o_i = \varphi(W_o V_i + U_o h_{i-1}). \quad (7)
\]

Here, \(c_i\) represents the memory unit of LSTM, then the memory unit \(c_i\) of unit time \(i\) in LSTM is adjusted to the new content \(c_i\) after input gate \(I_i\) and forgetting gate \(f_i\). The memory unit of LSTM can be calculated using equations (8) and (9).

\[
c_i = f_i c_{i-1} - I_i c_i, \quad (8)
\]

\[
c_i = \text{tanh}(W_c V_i + U_c h_{i-1}). \quad (9)
\]

Input gate \(I_i\) and forgetting gate \(f_i\) control the input of new content and the forgetting of old content, respectively. Here \(I_i\) and \(f_i\) can be calculated using equations (10) and (11), respectively.

\[
I_i = \varphi(W_i V_i + U_i h_{i-1}), \quad (10)
\]

\[
f_i = \varphi(W_f V_i + U_f h_{i-1}). \quad (11)
\]

At this point, when \(W\) and \(U\) are determined, the constructor \(F\) is uniquely determined. In LSTM, the optimization function \(\text{RMSProp}\) is usually introduced to determine \(W\) and \(U\).

A batch of samples with a capacity of \(N\) is randomly selected from the training set \([V_1, V_2, \ldots, V_N]\):

\[
g = \frac{1}{N} \sum_i (F(V_i; \theta) - m_i)^2, \quad (12)
\]

\[
r = pr + (i - p) g + g. \quad (13)
\]

Then calculate the updated parameters according to \(r\) and updated \(\theta\), and these can be calculated using equations (13) and (14).
\[ \Delta \theta = \frac{e}{\delta + \sqrt{r}} + g, \quad (13) \]
\[ \theta = \theta + \Delta \theta, \quad (14) \]
where \( e \) is the learning rate.

Because the music track is usually located in the middle of the two music units MFCC. Therefore, remove the fore and aft part of the two units of audio, respectively, in the collection of Sh and St, and the rest of the music as the main body in the collection of Sb. In music synthesis, a unit audio \( m1 \) is randomly selected from the set Sh as the input and H as the output, and then the output H synthesized by the algorithm is continuously matched with the unit music vector in S. The similarity matching strategy adopted in this article is Euclidean distance calculation. The nearest but music vector is the next unit music \( m1 + 1 \) predicted by the model, as shown in equations (15) and (16).

\[ m_{i+1} = S[x], \quad (15) \]
\[ x = \text{index} (\text{min} [d(h, m1), d(h, m2), \ldots, d(h, m_N)]), \quad (16) \]
where \( x \) is the index of unit music in data set \( S \).

The Euclidean distance \( d_{ab} \) between two audio units \( ma \) and \( mb \) is calculated as

\[ d_{ab} = \left( \sum_{j=1}^{n} (v_{a,j} - v_{b,j})^{2} \right)^{1/2}, \quad (17) \]
where \( j \) represents the \( j^{th} \) vector value of \( n \)-dimensional vector \( V \) of unit music \( m \).

This article first describes the relevant basic theory of music, including the basic properties of music, rhythm and beat, interval, national mode, and melody form. This is the foundation of music composition that needs to master, and then introduces the relevant knowledge of MIDI, the music standard format used in the data processing.

4. Experimental Results and Analysis

The data for this study is mostly obtained from an independently gathered collection of MIDI folk tunes. After preprocessing, this is submitted to our network for training. It is vital to train the system after it has been constructed. In the procedure of network design development, initially, the starting model’s parameters must be selected, and the choice of diverse variables has a large impact on the network’s integration. The values of the up-sampled pattern are standardized as inputs, and the dataset of MIDI training is up-sampled [24].

The seven criteria mentioned are taken as melody characteristic parameters: pitch range, pitch stability, pitch distribution, transverse dissonance interval ratio, vocal progression, progressive ratio, and vocal dissonance. For the results generated by the above melody fitting curves, the Euclidean distance of the melody characteristic parameters is calculated, and the results are shown in Table 1.

Using polynomial fitting, the components of the melody, as well as the rhythm of the feature parameters and the unique song, are obtained from the above results. When the melody of seven melody parameters of least Euclidean distance was compared with other approaches, this work discovered that the authentic track duo Euclidean distance is the closest to our aim. The comprehensive evaluation result \( p \) is also higher than other methods. The pitch stability of a two-part melody obtained after Gaussian function fitting is too high, beyond the range of (0, 20), with large pitch fluctuations, which is not suitable for the creation of a two-part melody. In general, polynomial curve fitting results are closer to the real form of the melody of the second part of the original song, the fitting effect is the best, and it is more suitable for the creation of the melody of the second part.

For the first time, artificial intelligence becomes employed in this work to transform song composition data into feature extraction and presentation. It was used for further feature extraction and had a considerable impact. This method extracts the majority of the capabilities of MIDI files. Artists use MIDI files to offer and save tune information in a report format that contains tune characteristics as well as more detailed data. The extracted MIDI music feature matrix is shown in Table 2:

In Table 2, note features are extracted from MIDI files. Each behavior has a note vector, and information such as instrument number, pitch, start time, end time, strength, speed, metronomic mark, and key of musical notes are recorded. As the MIDI music stored is in different modes, the extraction of MIDI music information in a report format that contains tune characteristics as well as more detailed data. The extracted MIDI music feature matrix is shown in Table 2:

In Table 2, note features are extracted from MIDI files. Each behavior has a note vector, and information such as instrument number, pitch, start time, end time, strength, speed, metronomic mark, and key of musical notes are recorded. As the MIDI music stored is in different modes, considering the modes of different music are different, the network training and composition generation effect will be affected. The next step is to carry out a unified mode conversion of MIDI music.

The network model will get more complicated if there are too many neurons in each layer, and the training will take up more memory. The ideal model parameters were identified, and different hidden layers and the number of units at each layer were set and tested numerous times to maximize training speed and minimize memory resource loss.

The following are the outcomes of the experiment:

(i) When the hidden layer has only one layer of Bi-GRU, the effect of the diverse number of neurons on correctness (accuracy) is shown in Table 3. According to this table, the numbers of neurons are 128, 256, and 512. The accuracy of the training set is 83.68%, 85.67%, and 96.89%. Similarly, the test accuracies are 79.68%, 82.68%, and 82.77%.

(ii) When the hidden layer has two Bi-GRU layers, the first layer contains 512 neurons, and the impact of the diverse numbers of neurons on the accuracy of the second layer is revealed in Table 4. According to this table, the numbers of neurons are 128, 256, and 512. The accuracy of the training set is 96.31%, 97.75%, and 98.02%. Similarly, the test accuracies are 85.62%, 86.61%, and 86.16%.

(iii) When the hidden layer has three Bi-GRU layers, the first layer has 512 neurons and the second layer has
The impact of the diverse number of neurons on the accuracy of the third layer is revealed in Table 5. According to this table, the numbers of neurons are 128, 256, and 512. The accuracy of the training set is 95.86%, 98.31%, and 99.32%. Similarly, the test accuracies are 87.15%, 88.73%, and 89.68%.

According to the experimental data, the model of an algorithm developed in this study has the best correctness and the lowest loss rate when associated with the other 3 models. This can be demonstrated that the model of algorithm proposed in this study outperforms the training feature of set depiction and network layout scheme, and is thus appropriate for learning the inner structural characters of sequences of note from the exercise set. The chromatographic distribution of the vector of the created music projected by dissimilar algorithm techniques is studied in this research to properly associate the music effect provided by every method.

Figure 3 shows the evaluation of the loss rate and accuracy of the algorithm design. This figure shows that basic_rnn, mono_rnn, and attention_rnn have the accuracy of 83%, 68%, and 73%, respectively. Similarly, their loss rate is 0.23, 0.53, and 0.44, respectively. While the accuracy of our proposed method is much higher than others, which is 99% and it has the smallest loss rate such as 0.03.

The chromatographic vector dispersion between the test set music and the music predicted in this work is depicted in Figure 4.

Among the 30 songs randomly selected in this experiment, there are 10 folk songs created by manual composers, 10 songs composed by the design model in this article, and 10 songs composed by the basic_RNN model in the Magenta project of Google laboratory. As the five parameters are equally important, the score results of all staff on these music indicators were averaged first to get the average score of the five parameters and then compared, respectively. Figure 4 shows the score comparison results of five indicators between the model composition in this article and the Magenta Project basic_RNN algorithm model composition. Figure 5 shows the scoring comparison results of five indicators between model composition and manual composition in this article (Figure 6).

According to the comparison results of indicators in the aforementioned figures, the sweetness and concordance of the music produced by the approach of this work and the music produced by the basic_RNN algorithm model are the same as those generated by manual composition, indicating that the music can meet people’s needs from the sensory point of view. From the perspective of style clarity, the music produced by the approach of this work is higher than that generated by the basic_RNN algorithm model. Because the basic_RNN algorithm model uses the basic one-hot coding to extract melody features as the neural network input, the encoding method of up-down sampling is adopted in this article to integrate pitch and time value. The rhythm sequence is generated while the pitch is generated, which makes the generated music retain the rhythm style.
The basic RNN algorithm model is more suitable for western music generation. In terms of the degree of innovation of the generated music, the network designed in this article is composed of a variety of algorithms. The bidirectional recurrent neural network learns the characteristics of the music from the training samples, and the motivational melody generated by the Markov model is the initial input sequence, which also introduces certain randomness and flexibility. This allows the output to be neither too random nor mechanical. However, from the perspective of hierarchy, there is a big gap between the music produced by the approach of this research work and the artificial composition, and the hierarchy of music will also affect the quality of music. The model designed in this article needs further improvement.

Table 5: Impact of the diverse number of neurons on correctness in third hidden layers.

| Number of neurons | 128 | 256 | 512 |
|-------------------|-----|-----|-----|
| Accuracy of training set (%) | 95.86 | 98.31 | 99.32 |
| Test set accuracy (%) | 87.15 | 88.73 | 88.75 |

Figure 3: Accuracy and loss rate comparison.

Figure 4: Experimental results of comparison methods. (a) The training sample, (b) basic_rnn, (c) mono_rnn, and (d) attention_rnn.
5. Conclusions

In conclusion, based on the music audio composed for operation objects in AI, this article draws lessons from the speech signal processing approach by utilizing MFCC as a characteristic vector. Here, the music tracks are represented as a music fragment sequence with time series characteristics, and the LSTM-RNN produced training is used as the training model. In addition, based on the music generator constructed by the deep learning model LSTM, this article also generates music of a specific music style by achieving 99% accuracy in different test schemes. According to the evaluation of test results, this model can perform music creation well, thus providing strong support for the feasibility of more music creation based on artificial intelligence algorithms in the future. So far, this article has realized music generation is based on a single type of music style. In the future, I will try to compose music in real-time according to the machine in different moods of users.

Data Availability

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

Conflicts of Interest

The author declares that there are no conflicts of interest or personal relationships that could have appeared to influence the work reported in this article.

References

[1] E. Loubet, C. Roads, and B. Robindore, “The beginnings of electronic music in Japan, with a focus on the NHK studio: the 1950s and 1960s,” Computer Music Journal, vol. 21, no. 4, 11 pages, 1997.
[2] X. Xia and J. Yan, “Construction of music teaching evaluation model based on weighted naïve bayes,” Scientific Programming, vol. 2021, Article ID 7196197, 9 pages, 2021.
[3] L. Yu, W. Zhang, J. Wang, and Y. Yu, “Sequence generative adversarial nets with policy gradient,” in Proceedings of the AAAI Conference on Artificial Intelligence, San Fransico, CA, USA, February 2017.
[4] T. N. Sainath, B. Kingsbury, and A. Mohamed, “Improvements to deep convolutional neural networks for LVCSR,” in Proceedings of the 2013 IEEE workshop on automatic speech recognition and understanding, pp. 315–320, IEEE, Olomouc, Czech Republic, December 2013.
[5] M. W. Browne, “Cross-validation methods,” Journal of Mathematical Psychology, vol. 44, no. 1, pp. 108–132, 2000.

[6] M. Akbari and J. Liang, “Semi-recurrent CNN-based VAE-GAN for sequential data generation,” in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2321–2325, Calgary, Alberta, April 2018.

[7] Y. Wang, “Music composition and emotion recognition using big data technology and neural network algorithm,” Computational Intelligence and Neuroscience, vol. 2021, Article ID 5398922, 11 pages, 2021.

[8] S. Sigita, E. Benetos, and N. Boulanger-Lewandowski, “A hybrid recurrent neural network for music transcription,” in Proceedings of the 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 2061–2065, IEEE, Brisbane, Australia, April 2015.

[9] J. A. Franklin, “Recurrent neural networks for music computation,” INFORMS Journal on Computing, vol. 18, no. 3, pp. 321–338, 2006.

[10] P. Dhanalakshmi, S. Palanivel, and V. Ramalingam, “Classification of audio signals using SVM and RBFNN,” Expert Systems with Applications, vol. 36, no. 3, pp. 6069–6075, 2009.

[11] B. Mathieu, S. Essid, and T. Fillon, “YAAFE, an easy to use and efficient audio feature extraction software,” ISMIR, vol. 2010, pp. 441–446, 2010.

[12] H. A. García Salas, A. Gelbukh, and H. Calvo, “Music composition based on linguistic approach,” in Proceedings of the Mexican International Conference on Artificial Intelligence, pp. 117–128, Springer, Pachuca, Mexico, November 2010.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in Neural Information Processing Systems, vol. 25, 2012.

[14] Y. Xu, “Systematic study on expression of vocal music and science of human body noise based on wireless sensor node,” Mobile Information Systems, vol. 2021, Article ID 9993019, 9 pages, 2021.

[15] T. Mikolov, A. Deoras, and D. Povey, “Strategies for training large scale neural network language models,” in Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, pp. 196–201, IEEE, Waikoloa, HI, USA, December 2011.

[16] N. Jaques, S. Gu, R. E. Turner, and E. Douglas, “Tuning recurrent neural networks with reinforcement learning,” in Proceedings of the ICLR Conference, Toulon, France, April 2017.

[17] R. Cowie and R. R. Cornelius, “Describing the emotional states that are expressed in speech,” Speech Communication, vol. 40, no. 1-2, pp. 5–32, 2003.

[18] S. Bonnabel, “Stochastic gradient descent on riemannian manifolds,” IEEE Transactions on Automatic Control, vol. 58, no. 9, pp. 2217–2229, 2013.

[19] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] F. A. Gers, J. Schmidhuber, and F. Cummins, “Learning to forget: continual prediction with LSTM,” Neural Computation, vol. 12, no. 10, pp. 2451–2471, 2000.

[21] H. Muaidi, A. Al-Ahmad, T. Khdoor, S. Alqrainy, and M. Alkoffash, “Arabic audio news retrieval system using dependent speaker mode, Mel frequency cepstral coefficient and dynamic time warping techniques,” Research Journal of Applied Sciences, Engineering and Technology, vol. 7, no. 24, pp. 5082–5097, 2014.