Hindawi
Security and Communication Networks
Volume 2023, Article ID 9850239, 1 page
https://doi.org/10.1155/2023/9850239

Retraction

Retracted: Note Detection in Music Teaching Based on Intelligent Bidirectional Recurrent Neural Network

Security and Communication Networks

Received 1 August 2023; Accepted 1 August 2023; Published 2 August 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Y. Yue, "Note Detection in Music Teaching Based on Intelligent Bidirectional Recurrent Neural Network," *Security and Communication Networks*, vol. 2022, Article ID 8135583, 9 pages, 2022.

Hindawi
Security and Communication Networks
Volume 2022, Article ID 8135583, 9 pages
https://doi.org/10.1155/2022/8135583

Research Article

Note Detection in Music Teaching Based on Intelligent Bidirectional Recurrent Neural Network

Ya Yue

College of Music and Dance, Liaocheng University, Liaocheng 252000, Shandong, China

Correspondence should be addressed to Ya Yue; yueya@lcu.edu.cn

Received 22 March 2022; Revised 23 April 2022; Accepted 28 April 2022; Published 18 June 2022

Academic Editor: Muhammad Arif

Music education is an essential and significant link in a quality education, as it can assist pupils improve their integrity and nurture noble character. The evident distinction between music teaching and teaching in other disciplines is that music teaching can provide aesthetic education to students in order to improve students' self-cultivation and overall temperament and basically play a role in developing people in a holistic fashion. Note detection is an important content in music teaching. Instrument tuning, computerized score recognition, music database search, and electronic music synthesis all benefit greatly from note detection technologies. In note detection, there are problems such as difficult one-to-one correspondence between estimated pitches and standard frequencies, a narrow range of identifiable pitches, poor robustness of the recognition process, and low recognition rate. In this context, this work proposes an automatic note detection in music teaching based on deep learning. It uses a convolutional neural network (CNN) and a bidirectional long-short-term memory (BiLSTM) network to build a deep neural network model, called convolutional neural network Bidirectional Long Short-Term Memory (CNN-BiLSTM), using this network to conduct in-depth research on note detection. First, based on the current research status, a deep neural network model based on CNN and BiLSTM is proposed to detect musical notes. The network can independently mine and learn the deep-level features of music signals and has better feature extraction and generalization capabilities. Second, the experimental results are evaluated using different evaluation metrics. Experiments show the network model can significantly improve detection accuracy and can efficiently detect notes in music teaching.

1. Introduction

The speed of contemporary social development is accelerating, which has led to the continuous exploration and practice of education. In today's 21st century, compulsory education assumes the main responsibility of cultivating new talents and plays an irreplaceable role in the process of improving the quality of people. In the current social development, the key to the success of compulsory education lies in reform and innovation. Therefore, since the beginning of the last century, a new round of basic education curriculum reform has been launched. After that, more and more people began to pay attention to improvement for students' comprehensive ability in all aspects; thus, the concept of quality education began to spread. With its particularity, the music discipline gradually occupies an increasingly important position in quality education. For quality education, music education is an indispensable and important link, which can help students to improve their integrity and cultivate noble character. The evident distinction between music teaching and teaching in other disciplines is that music teaching can provide aesthetic education to students in order to improve students' self-cultivation and overall temperament and basically play a role in developing people in a holistic fashion. This teaching goal is deeply in line with the concept of humanistic teaching concept. The practice has shown that learning music can help students improve their reading, creativity, and imagination skills. Students can improve their language skills and broaden their horizons in the process of appreciating music. In the process of music learning, students can improve their imagination and innovation ability by creating music. In the

process of appreciating many musical works, students can experience the musical ideas expressed by the musicians, and at the same time, they can have an ideological resonance with the musicians. In addition, students can improve their music performance level and self-confidence in the process of music performance, which is beneficial to their future development [1–5].

In the context of today's era, the development of science and prosperity of culture and art has brought about the integration of science and technology and culture and has achieved the prosperity of cultural development. The new discipline of music technology is the product of the combination of music and technology. The music information retrieval technology emerging is an important part of music technology. Note detection and recognition is an important branch of music information retrieval. Note detection is an important research content in the field of music signal analysis and processing. Instrument tuning, computerized score recognition, music database search, and electronic music synthesis all benefit greatly from note recognition technology. The growth of music technology and new electronic businesses is greatly aided by the development of note detection [6–10].

The value of musical scores is to music as the value of ancient books is to history. With the rapid development of the Internet, people have more and more convenient ways to obtain musical scores. People can share their favorite or collected musical scores online for everyone to discuss and appreciate together. People are eager to have such a technology that can recognize the melody or voice they hear and improve the current environment. Music recognition technology first appeared in the 1970s, the complete note detection, the recognition system was in the 1980s, and in the 20th century, a complete piano rhythm has been developed [11–15].

With the transformative development of computer technology, computational intelligence methods can be combined with various disciplines. In particular, the development of neural networks has facilitated the solution of various intelligent tasks. Based on a neural network, this work is committed to building a method for automatic detection and recognition of musical notes in the process of music teaching, so as to improve the efficiency and quality of subject music teaching.

The structure of this article is organized as follows. The literary works related to this study are presented in Section 2. The method of the proposed work is explained in Section 3. The experimentation and results of the suggested method are presented in Section 4. Finally, Section 5 summarizes the paper's main points.

2. Related Work

The evaluation of pitch or note is the first step in automatic score recognition. Due to the widespread use of pitch estimation in speech recognition and general audio categorization and analysis, there are several scholarly articles on the issue. In [16], a pitch estimation method based on a recurrent neural network was reported. Because of the shorter time range of this technique, it may now be used to estimate pitch on signals with quick pitch alterations, but it comes at the cost of increased computer complexity. Automatic note recognition for music saved in stereo or monophonic form is performed by [17] using an adaptive template matching method, and their system is able to identify the musical instrument that plays the note. Reference [18] developed a generalization spectrum-based pitch detection and estimation technique. Traditional autocorrelation and cepstrum approaches were also compared. They found that their method outperformed the traditional approach to pitch identification, especially at low signal-to-noise ratios, but that its estimation accuracy lagged behind that of the traditional method. For both popular and classical music, a novel pitch estimation approach has been proposed by [19].

Articles on beat or rhythm recognition have increased in recent years. Reference [20] proposed a beat tracking algorithm for music without percussion instruments. The method can not only detect beat information spaced any more than a quarter note but also analyze the structure of music beats on a larger time scale. In this way, the change points of advanced musical structures such as harmony and rotation can be found. Reference [21] used a wavelet analysis technique based on linear phase labor transform to analyze the rhythm information of music. He decomposed the possible rhythm signal into several components and then analyzed it through the phase consistency constraint. His method is more effective for musical rhythms expressed by percussion instruments. Paper [22] proposed an adaptive learning method based on maximum a posteriori estimation for tempo and analysis in music signals. The experimental results show that the method is relatively stable and not very sensitive to the size of the analysis window. Study [23] analyzed the structure of dance movements by detecting the rhythm of accompanying music.

Melody and harmony are significant high-level information in musical compositions, but extracting them from music records based on generic audio signals is difficult, and there are few literature publications on the subject. Based on wavelet transform and self-organizing neural network, [24] developed an automatic recognition approach for multitimbral harmony. Without knowing specific instrument timbres or note sequences, their system can automatically classify harmonic audio clips. Paper [25] described numerous approaches for extracting the primary melody track from MIDI files automatically. A method for recognizing harmony was proposed in [26]. The method can be divided into two stages, and the set of possible harmony candidates is given in the local recognition stage. Then some global rules were used to find the most suitable harmony as the final result. However, how to find or represent the main theme from general WAVE files is a difficult problem.

Under the influence of instrument reverb, it is nearly impossible to make out individual musical notations. However, in actual musical compositions, there are frequently many polyphonic elements. There is not much literature on polyphonic score recognition, but research in this area has increased in the last two years. Reference [27]

proposed a method for analyzing polyphonic music scores based on a dynamic Bayesian network. Their approach, which emphasizes the modeling of sound production processes, enables the tracking of beat and pitch trajectories of polyphonic music. Study [28] used multiresolution Fourier transform coefficients to identify the musical score of polyphonic music played by the piano. The experimental results show that better pitch detection results can be obtained under the condition that some restrictions are imposed on the performance. Scholars of [29] studied how to use stereo music signals for polyphonic music score recognition. They separated distinct sound sources by comparing the ratio of the signal levels of the two channels to conduct score recognition for trios. Reference [30] developed an overtone tracking technique for automatic note recognition in polyphonic piano music based on a combination of an auditory model and an adaptive oscillator network.

3. Method

In this chapter, we define the CNN, BiLSTM, music teaching data preprocessing, and music note detection with CNN-BiLSTM in detail.

3.1. Convolutional Neural Network. The main structure of CNN includes a convolutional layer, pooling layer, and fully connected layer. CNN is a type of deep neural network, and many variants have also appeared in recent years. CNN usually includes multilayer convolutional layers and pooling layers, the pooling layer is generally after the convolutional layer, and the result obtained after learning is an image-specific feature vector space. The fully connected layer plays the role of a classifier, which maps the information mapped to the feature space to the label space of the sample through the fully connected calculation to complete the recognition work, and finally transmits the result to the output layer.

The convolutional layer mainly extracts data features, replacing the method of manually designing data features in traditional computer vision methods, and automatically extracts abstract features that CNN can recognize through sliding calculation of convolution kernels in the graph. Generally speaking, the first convolutional layer in CNN can only extract relatively low-level features such as edges, lines, and angles. The deeper the network layer, the more abstract the feature extraction of the convolutional layer. In CNN, the convolution kernel of the convolution layer performs window sliding scanning on the image, and the final result is the local feature map extracted by the convolution layer, also known as the local receptive field. The calculation formula is

$$Q = \sum_{l=1}^{ab} w_l u_l. \tag{1}$$

The local receptive field is also one of the most advantageous features of convolutional layers. Similar to the way humans perceive external things, each neuron does not need to perceive global information but only needs to perceive local areas. The local information obtained from the bottom layer is then synthesized at a higher level to obtain global information. The weight-sharing mechanism is another advantage of the convolutional layer in CNN, which is mainly manifested in that when different convolution kernels traverse the entire image, the weights are fixed. Therefore, CNN avoids the complex work of learning new weights and biases updated by each neuron in the hidden layer and only needs to learn a set of weights and a single bias for global feature extraction.

The essence of pooling is a downsampling operation. Pooling layers are usually followed by convolutional layers and reduce the dimension of features obtained through convolutional layer operations. The principle used by the pooling layer is mainly the local correlation of the image, and its main operation mode is to take the maximum value or average value from the target area to subsample the image. This not only preserves useful feature information but also improves the robustness of the system. Even if the input data to the previous network had an acceptable bias, pooling would still return the same results as would be obtained with unbiased data. Pooling can greatly reduce the risk of model overfitting and improve the accuracy and fault tolerance of the model.

The pooling layer can integrate the information of the feature map, reduce the size of the feature map, and reduce the computational cost of the model. There are two most commonly used pooling methods: average pooling and max pooling. The method of average pooling is to average the values in the pooled kernel to obtain new eigenvalues, while the maximum pooling is to extract the largest value combination output in the pooling kernel. Average pooling can maximize the preservation of background information. The maximum pooling can effectively extract the feature texture, reduce the impact of useless information on the experimental results, and enable the model to achieve better recognition results under the premise of effectively reducing the data dimension. The calculation is

$$Y_k^C = g_p\left(X_k^C\right). \tag{2}$$

The activation function makes a nonlinear change to the feature map through a nonlinear function, which not only acts as a decision function but also promotes the network to learn more complex nonlinear mapping. In addition, by choosing an appropriate activation function, the training process of the network can be accelerated. For the feature map, the activation function g calculates it element by element to get the result. At present, the more commonly used activation functions of CNNs are Sigmoid, Tanh, and ReLU:

$$\begin{aligned} \text{Sigmoid} &= \frac{1}{(1+\exp(-x))}, \\ \text{Tan}h &= \frac{(\exp(x)-\exp(-x))}{(\exp(x)+\exp(-x))}, \\ \text{ReLU} &= \max(x, 0). \end{aligned} \tag{3}$$

Compared with the convolutional layer and the pooling layer, the biggest feature of the fully connected layer is that all units in each layer are connected to the neurons in the

previous layer. This enables the comprehensive processing of features extracted by convolutional and pooling layers. Usually, layers other than the input and output layers are considered hidden layers.

During the training process of the neural network, for the same neuron or convolutional layer, the input data of different batches may have a large difference in distribution. This phenomenon is called the internal covariate shift in the neural network. Due to the existence of internal covariate offset, a small learning rate can only be selected during network training to avoid the resulting oscillation, resulting in slow network convergence. Batch normalization solves the problem of network training difficulties caused by internal covariate shifts by normalizing the input tensors of a batch. Specifically, the normalization method adopted by batch normalization is to subtract the mean of a batch tensor and divide it by the standard deviation to convert it into distribution with a mean of 0 and a variance of 1.

3.2. Bidirectional Long Short-Term Memory. BiLSTM is a new type of network that has been widely employed in various fields and has been upgraded on the basis of traditional recurrent neural network (RNN). People apply the concept that prior experience and memory affect human cognition in neural networks, based on the inspiration provided by the human brain, and the properties of RNN also highlight this element. RNN is a relatively special neural network structure. The biggest difference between it and CNN is that it not only needs to process the current input data signal but also has the function of memory; that is, what RNN needs to process is the new data generated by the combination of the present and the previous.

The reason why RNN is called a recurrent neural network is that it has a special memory function, which enables RNN to learn both before and after information at the same time, so as to achieve more perfect feature extraction. The schematic diagram of the hidden layer expansion of RNN is shown in Figure 1, where $t-1$, t, $t+1$ represent time series. X represents the input sample data, S_t is the memory of the input sample at time t, W, represents the weight value in the hidden layer, B represents the weight value of the input sample at this moment, and V represents the output sample weight value after network learning. It should be noted that the weight parameters will be continuously updated during the calculation process.

Despite the fact that RNNs are now widely employed and have produced positive results, they still have some flaws. The key issues are gradient disappearance and gradient explosion of RNN during training; these issues make RNN training challenging, and the application is constrained in many ways. The long-short-term memory (LSTM) network was born as a result of this, and it used gating units and memory techniques to greatly ease the gradient problem. The key part of LSTM is its cell state update mechanism and the gate structure. In the LSTM network, the addition and deletion of information are mainly realized through the three gate structures forget gate, memory gate, and output gate. This gate structure identifies which information should be kept and which information should be discarded during the training of the network.

In LSTM, we think that information is transmitted through the cell state. The content of its transmission can be regarded as the memory of the network, and its essence is the useful feature information extracted by the previous nerve cells. When ignoring other factors, LSTM can transmit useful information obtained during sequence processing in the network through cellular memory. Therefore, even the feature information that is trained in the early access network can be carried to the later cells, and this function makes up for the short-term memory of the neural network. The learning process of LSTM is

$$\begin{aligned} f_t &= \phi\left(W_f\left[h_{t-1}; x_t\right] + b_f\right), \\ i_t &= \phi\left(W_i\left[h_{t-1}; x_t\right] + b_i\right), \\ o_t &= \phi\left(W_o\left[h_{t-1}; x_t\right] + b_o\right), \\ h_i &= o_t * \tanh\left(c_t\right), \\ c_t &= f_t * c_{t-1} + i_t * \tanh\left(W_s\left[h_{t-1}; x_t\right] + b_s\right). \end{aligned} \tag{4}$$

The three gate structures of forget gate, memory gate, and output gate enable the network to remember features from earlier stages of the sequence. This enables LSTMs to store, read, and update historical information over long distances, capturing long-term dependencies between data.

But one disadvantage of traditional LSTM is that its processing of information is limited to forward and backward. However, the environment in which the neural network works is complex and changeable, and sometimes just looking at the previous information is not enough, so a bidirectional neural network BiLSTM is proposed. BiLSTM can achieve the purpose of combining contextual information by processing data in two directions at the same time through two separate hidden layers. This innovation makes up for the shortcomings of LSTM. BiLSTM is composed of a forward LSTM and a backward LSTM, and its network structure is shown in Figure 2.

Two LSTMs in opposite directions exist simultaneously in a BiLSTM model. The relevant calculations for LSTMs have been described above. If it is considered that at time t, the hidden state of the forward LSTM output is h_{t0} and the hidden state of the reverse LSTM output is h_{t1}, then the output of the hidden state of BiLSTM at this time is

$$h_t = h_{t1} * h_{t2}. \tag{5}$$

BiLSTM's bidirectional structure allows it to better handle data's pre- and postdependency and mine the discriminative deep-level characteristics in the original data. This model can also successfully prevent the above-mentioned difficulties of gradient disappearance and gradient explosion.

3.3. Music Teaching Data Preprocessing. Often, data preprocessing is required in the early stages of research work. The preprocessing of music data usually includes three parts: noise processing, music data interception, and data standardization. Noise processing is mainly to remove all kinds

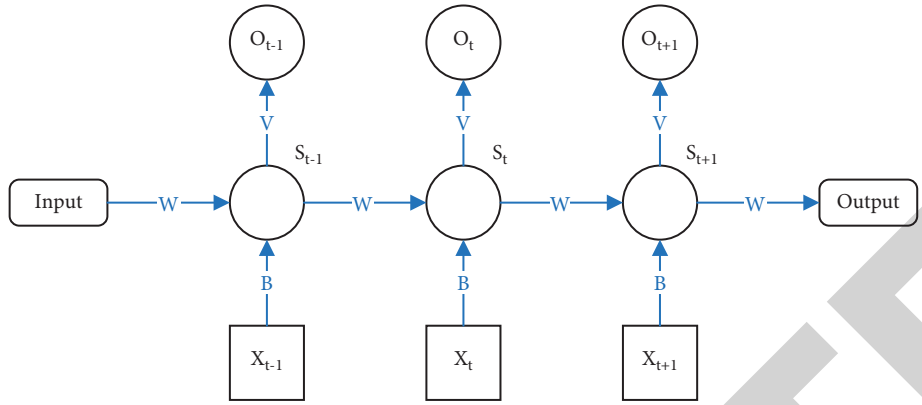

Figure 1: The structure of the RNN hidden layer.

of noise in the music signal, better retain the useful information of the music signal, and prevent the noise from adversely affecting the note detection accuracy. Music data are intercepted. Interception refers to segmenting the original data and training using music snippets to improve the network's computation speed. The mapping of signals from various ranges to the intended range is known as data normalization.

Music signals are extremely sensitive to noise, and in the process of acquiring music signals, the signals may be affected by various external factors and generate noise during the acquisition process. The three kinds of noises in music signals are mainly power frequency interference, base point interference, and baseline drift. These noises increase the difficulty of note detection in music signals. Therefore, removing the noise interference of music signals is of great significance to improve the accuracy of note detection. In this paper, a 40-order band-pass filter is designed to filter the original music signal. Band-pass filtering is a filtering method that only allows certain frequencies to pass, while effectively suppressing the signals of other frequencies.

Deep learning can mine the internal hidden information from the music signal and realize the automatic extraction of features. This method can not only enhance the generalization ability of the network but also maximize the use of the features learned by the network to avoid distortion of the input data. However, the music signal generally has a large data length, and the length varies. Therefore, in the study of note detection, it is necessary to segment the music signal. Using segmented music clips for research can improve computational speed and reduce computational complexity. This work divides music data into fixed-length segments for network training and testing.

In the process of music signal acquisition, there is a certain relationship between the music data of the same individual in different periods; that is, the distance between the two will continue to change with the passage of time, resulting in amplitude scaling and offset, resulting in errors in the experimental results. Therefore, in order to achieve a better experimental effect, the music signal needs to be normalized to reduce the amplitude difference between the music signals. The normalization process for each music signal sample is given by

$$x' = \frac{(x - x_{\min})}{(x_{\max} - x_{\min})}. \tag{6}$$

Normalizing the input data can convert the amplitudes of all samples between 0 and 1, which enhances the stability of the data and facilitates the learning and training of the network.

3.4. Music Note Detection with CNN-BiLSTM. CNN has unique advantages in data feature extraction, but it is difficult to deal with the dependencies between data. However, BiLSTM can learn data features in the time dimension, which makes up for the shortcomings of CNN. Therefore, this paper considers combining the two neural networks to explore whether higher note detection accuracy can be obtained. In this chapter, a network model combining CNN and BiLSTM is proposed to achieve high-precision detection of musical notes.

In this paper, the composite model built on the basis of the combination of CNN and BiLSTM is named CNN-BiLSTM. The model can directly perform feature extraction on the preprocessed original music signal, and no complicated manual feature extraction process is needed. The model can also adaptively explore and learn the deep underlying structure and feature information of various music signals. The trained network can detect musical notes with high accuracy after testing. The suggested network structure is shown in Figure 3.

The starting part of the network structure is the input layer, and the network input of the training part is the music signal data. In order to fit the input of the CNN, it is processed into a two-dimensional data form. First, CNN is used to extract features from the data. The CNN used here mainly includes the convolutional layer and pooling layer. In the convolutional layer, the feature extraction work is realized by the convolution kernel through a sliding window on the input samples, and a new feature vector is formed by point-by-point multiplication. During the training process, the weights of the convolution kernels are continuously adjusted and updated to minimize the error and obtain valuable spatial feature information in the data. In order to reduce the size of the input data, a pooling layer with a stride

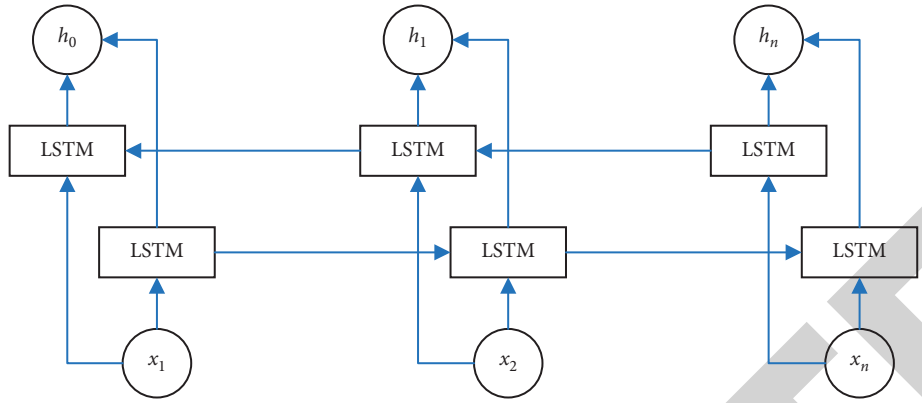

Figure 2: The structure of BiLSTM.

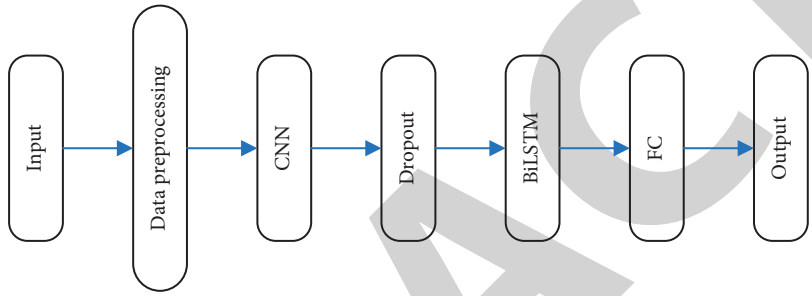

Figure 3: The structure of CNN-BiLSTM.

of 5 is used for filtering the feature maps after the first three convolutional layers. According to the data size output by the convolutional layer of the fourth layer, a pooling layer with stride 2 is selected for the pooling of the last layer. After the CNN is trained on the input data, the resulting network is put into BiLSTM for further feature extraction to obtain the before-and-after correlation of the data. The final stage of the network is the fully connected layer, which outputs the result of note detection.

The CNN model can convert the original input into a feature vector representation through convolution kernels, sliding windows, and pooling operations, so as to better capture the local features and deep features in the input signal. The source of the idea of the deep neural network proposed in this section is the powerful advantages of CNN in extracting spatial feature information, but it is difficult to learn the dependencies between input data, and there are some shortcomings in processing continuous time series signals. But BiLSTM can better understand the relationship between the context and makeup for the shortcomings of CNN in this regard. Therefore, this paper combines the advantages of both, proposes a CNN-BiLSTM deep learning model, and uses this network to achieve high-performance detection of musical notes.

4. Experiments and Results

4.1. Dataset and Details. The dataset used in this work is a self-made dataset, which contains a total of 89,471 music clips. Among them, 68,302 music clips constitute the training set, and 21,169 music clips constitute the test set. In this work, the evaluation metrics used are precision and recall. The experimental platform environment is illustrated in Table 1.

Table 1: Experiment environment.

Item	Name
Operating system	Ubuntu 16.04
CPU	Intel Core i7-6700
Memory	32 GB
Deep learning framework	PyTorch 1.6

4.2. Comparison with Other Note Detection Methods. First, this work is compared with other detection methods to verify the validity and correctness of the method designed in this paper. The methods involved include BP, RBF, and LSTM methods, and the experimental results are illustrated in Table 2.

Compared with other methods, the CNN-BiLSTM method proposed in this work can achieve 96.3% precision and 93.3% recall, which corresponds to the best performance. Compared with the best performing LSTM method in the table, this method can achieve 2.2% precision improvement and 1.4% recall improvement. This verifies the validity and correctness of the method designed in this work.

4.3. Effect of Different Learning Rates. The learning rate is a hyper-parameter in the neural network training process that has a significant impact on the experimental findings. The

Table 2: Comparison with other note detection methods.

Method	Precision	Recall
BP	87.50	85.20
RBF	91.20	86.70
LSTM	94.10	91.90
CNN-BiLSTM	96.30	93.30

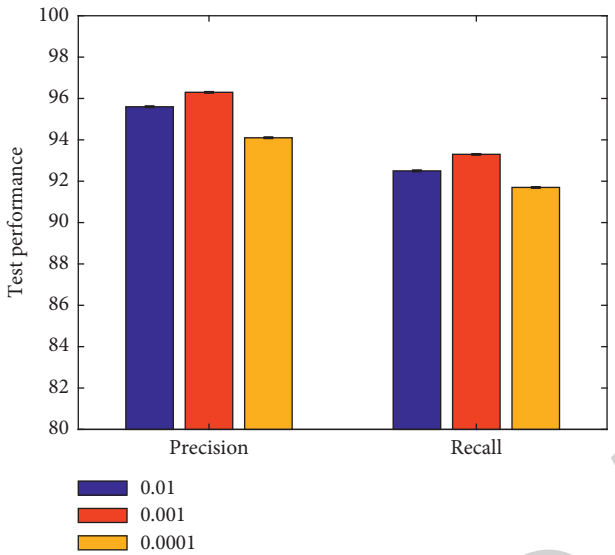

Figure 4: Effect of different learning rates.

speedy convergence of the objective function to the local minimum is aided by a proper learning rate setting. If the learning rate is set too low at the start, the network's training time for the target data will be extended, and the network's convergence speed will be slowed. On the other hand, if the initial learning rate is set too high, the network may fail to converge, resulting in oscillation. In this paper, under the same mathematics and network structure, different learning rates are set for comparative experiments. The specific results are illustrated in Figure 4.

It is obvious that note detection can achieve the best performance when the learning rate of the network is 0.01. When the learning rate decreases or increases, the precision and recall of the network both decrease to varying degrees. Therefore, in this work, the learning rate of the deep network is set to 0.001.

4.4. Effect of Different Optimizers. During backpropagation of neural networks, there are different optimizers to choose from. To evaluate the impact of different optimizers on the note detection performance, this work conducts comparative experiments to compare the detection accuracy when using the stochastic gradient descent (SGD) and Adam optimizers, respectively. The experimental results are shown in Figure 5.

It is obvious that training the neural network with the Adam optimizer achieves the best performance for note detection. Compared to using the SGD optimizer, using the Adam optimizer can achieve a 1.1% improvement in precision and 0.8% improvement in recall. Therefore, in this work, the latter is used to optimize the network.

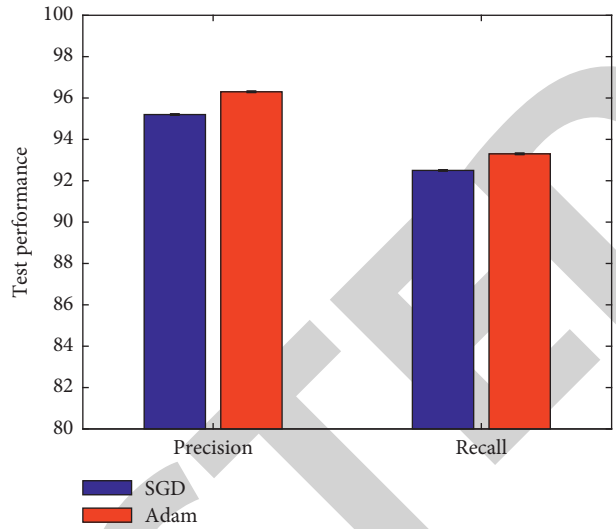

Figure 5: Effect of different optimizers.

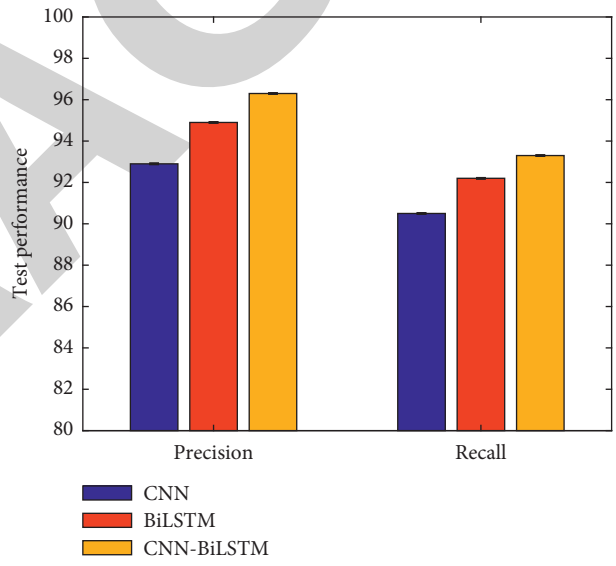

Figure 6: Effect of network combination.

4.5. Effect of Network Combination. As mentioned earlier, the CNN-BiLSTM network proposed in this work is composed of two modules, CNN and BiLSTM. To verify the effectiveness of this combined strategy, this work conducts comparative experiments to compare the detection performance of CNN-BiLSTM and two separate modules respectively. The experimental results are illustrated in Figure 6.

The experimental results prove that neither CNN alone nor BiLSTM alone can achieve the best performance. The highest precision and highest recall can be obtained on the dataset only after combining the two to build a CNN-BiLSTM network. Therefore, this further verifies the

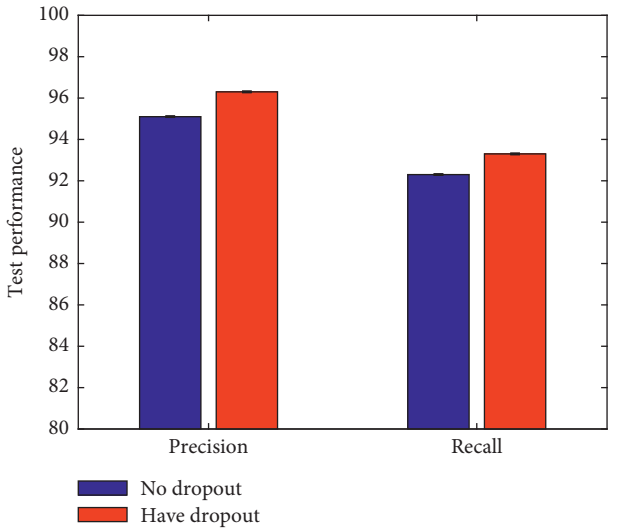

Figure 7: Effect of dropout strategy.

effectiveness and correctness of the CNN-BiLSTM network proposed in this work.

4.6. Effect of Dropout Strategy. The dropout strategy is used in the CNN-BiLSTM method suggested in this paper to avoid the network's overfitting problem. This work compares note detection performance without and with this technique to verify the effectiveness of this strategy. The experimental results are illustrated in Figure 7.

Obviously, compared with not using the dropout strategy, 1.2% precision improvement and 1.0% recall improvement can be obtained with this strategy. This verifies the correctness and reliability of the strategy used in this work.

5. Conclusion

The proposal of the music subject's core literacy is a specific necessity for teaching the music topic. As an important way of aesthetic education, music education can help students improve their aesthetic taste and deepen their emotional experience. As a part of the culture, music can reflect the connotation of different cultural regions. At the same time, the educational method can help pupils build comprehension skills such as creativity, cooperation, and imagination. In music education, note recognition is a crucial skill. Note detection technology has very important application value in musical instrument tuning, computer automatic score recognition, music database retrieval, and electronic music synthesis. This paper offers the CNN-BiLSTM, a deep neural network model for note identification in music education. The most notable benefit of this model is that it eliminates the need for a complex manual feature extraction approach in the early stages of detection work. It can automatically mine and learn the deep-level effective features of music data, have better feature extraction ability and generalization ability, and use the obtained data features to automatically detect notes. A comprehensive and systematic experiment is carried out in this work to verify the validity and correctness of the network designed in this work for note detection in music teaching.

Data Availability

The datasets used during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The author declares that he has no conflicts of interest.

References

[1] X. Wang and Y. Chen, "Music teaching platform based on FPGA and neural network," *Microprocessors and Microsystems*, vol. 80, Article ID 103337, 2021.

[2] C. P. Schmidt, "Systematic research in applied music instruction: a review of the literature," *Visions of Research in Music Education*, vol. 16, no. 1, p. 100, 2021.

[3] M. Müller, "An educational guide through the FMP notebooks for teaching and learning fundamentals of music processing," *Signals*, vol. 2, no. 2, pp. 245–285, 2021.

[4] W. Duan, J. Gu, M. Wen, G. Zhang, Y. Ji, and S. Mumtaz, "Emerging technologies for 5G-IoV networks: applications, trends and opportunities," *IEEE Network*, vol. 34, no. 5, pp. 283–289, 2020.

[5] Y. Chen, "Optimization of music teaching methods based on multimedia computer-aided technology," *Computer-Aided Design and Applications*, vol. 18, pp. 47–57, 2020.

[6] E. Shatri and G. Fazekas, "Optical Music Recognition: State of the Art and Major Challenges," 2020, https://arxiv.org/abs/2006.07885.

[7] S. Hizlisoy, S. Yildirim, and Z. Tufekci, "Music emotion recognition using convolutional long short term memory deep neural networks," *Engineering Science and Technology, an International Journal*, vol. 24, no. 3, pp. 760–767, 2021.

[8] Z. J. Calvo, J. Hajic, and A. Pacha, "Understanding optical music recognition," *ACM Computing Surveys*, vol. 53, no. 4, pp. 1–35, 2020.

[9] E. Brusa, C. Delprete, and L. G. Di Maggio, "Deep transfer learning for machine diagnosis: from sound and music recognition to bearing fault detection," *Applied Sciences*, vol. 11, no. 24, Article ID 11663, 2021.

[10] S. Visnu Dharsini, B. Balaji, K. S. Kirubha Hari, and Sridharshini, "Music recommendation system based on facial emotion recognition," *Journal of Computational and Theoretical Nanoscience*, vol. 17, no. 4, pp. 1662–1665, 2020.

[11] Y. Wang, A. Ren, M. Zhou, W. Wang, and X. Yang, "A novel detection and recognition method for continuous hand gesture using FMCW radar," *IEEE Access*, vol. 8, Article ID 167264, 2020.

[12] J. Grekow, "Music emotion recognition using recurrent neural networks and pretrained models," *Journal of Intelligent Information Systems*, vol. 57, no. 3, pp. 531–546, 2021.

[13] S. Edirisooriya, H. W. Dong, J. McAuley, and T. Berg-Kirkpatrick, "An Empirical Evaluation of End-To-End Polyphonic Optical Music Recognition," 2021, https://arxiv.org/pdf/2108.01769.

[14] R. Sarkar, S. Choudhury, S. Dutta, A. Roy, and S. K. Saha, "Recognition of emotion in music based on deep convolutional neural network," *Multimedia Tools and Applications*, vol. 79, no. 1-2, pp. 765–783, 2020.

[15] A. M. Proverbio, E. Camporeale, and A. Brusa, "Multimodal recognition of emotions in music and facial expressions," *Frontiers in Human Neuroscience*, vol. 14, p. 32, 2020.
[16] K. Han and D. Wang, "Neural network based pitch tracking in very noisy speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2158–2168, 2014.
[17] A. Rebelo, I. Fujinaga, F. Paszkiewicz, A. R. S. Marcal, C. Guedes, and J. S. Cardoso, "Optical music recognition: state-of-the-art and open issues," *International Journal of Multimedia Information Retrieval*, vol. 1, no. 3, pp. 173–190, 2012.
[18] W. J. Hess, "Pitch and voicing determination of speech with an extension toward music signals," *Springer Handbook of Speech Processing*, pp. 181–212, 2008.
[19] N. C. Maddage, "Automatic structure detection for popular music," *Ieee Multimedia*, vol. 13, no. 1, pp. 65–77, 2006.
[20] M. Goto and Y. Muraoka, "Real-time beat tracking for drumless audio signals: chord change detection for musical decisions," *Speech Communication*, vol. 27, no. 3-4, pp. 311–335, 1999.
[21] S. Nozaradan, "Exploring how musical rhythm entrains brain activity with electroencephalogram frequency-tagging," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1658, Article ID 20130393, 2014.
[22] J. Pei, K. Zhong, M. A. Jan, and J. Li, "Personalized federated learning framework for network traffic anomaly detection," *Computer Networks*, vol. 209, 2022.
[23] T. Shiratori and K. Ikeuchi, "Synthesis of dance performance based on analyses of human motion and music," *Information and Media Technologies*, vol. 3, no. 4, pp. 834–847, 2008.
[24] Y. Hu and G. Liu, "Instrument identification and pitch estimation in multi-timbre polyphonic musical signals based on probabilistic mixture model decomposition," *Journal of Intelligent Information Systems*, vol. 40, no. 1, pp. 141–158, 2013.
[25] J. Salamon and E. Gomez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio Speech and Language Processing*, vol. 20, no. 6, pp. 1759–1770, 2012.
[26] Y. R. Pandeya, B. Bhattarai, and J. Lee, "Music video emotion classification using slow-fast audio-video network and unsupervised feature representation," *Scientific Reports*, vol. 11, no. 1, Article ID 19834, 2021.
[27] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: an overview," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, 2019.
[28] F. Argenti, P. Nesi, and G. Pantaleo, "Automatic transcription of polyphonic music based on the constant-Q bispectral analysis," *IEEE Transactions on Audio Speech and Language Processing*, vol. 19, no. 6, pp. 1610–1630, 2011.
[29] Y.-T. Wu, B. Chen, and L. Su, "Multi-instrument automatic music transcription with self-attention-based instance segmentation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2796–2809, 2020.
[30] E. Nakamura, K. Yoshii, and S. Sagayama, "Rhythm transcription of polyphonic piano music based on merged-output HMM for multiple voices," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 794–806, 2017.