Hindawi
Mobile Information Systems
Volume 2021, Article ID 1738104, 7 pages
https://doi.org/10.1155/2021/1738104

Research Article

Research on Music Teaching and Creation Based on Deep Learning

Mingxing Liu

Shanxi Normal University, Linfen, Shanxi 041000, China

Correspondence should be addressed to Mingxing Liu; liumingxingyy@sxnu.edu.cn

Received 14 October 2021; Revised 5 November 2021; Accepted 11 November 2021; Published 2 December 2021

Academic Editor: Ateeq Rehman

Under the background of quality education, music learning is also changing, from the original shallow learning to deep learning gradually. In-depth learning is a new teaching concept, which pays full attention to students' perception and exploration of music so that students can fully experience the charm of music. It can not only help students master more music knowledge and improve their music skills but also cultivate students' music literacy and enhance their music ability (Świechowski, 2015). Therefore, in junior high school music teaching, teachers should actively apply the deep learning model and then improve the teaching level and comprehensively cultivate students' music literacy (Whitenack and Swanson, 2003). In this paper, two convolution-based deep learning models, Breath1d and Breath2d, were designed and constructed, and a multilayer perceptron (MLP) was used as a benchmark method for performance evaluation, and a long short-term memory (LSTM) network is applied for the classification task. This paper discusses the value and application strategies of deep learning in junior high school music teaching and hopes to provide some reference for all educational colleagues (Zhang and Nauman 2020).

1. Introduction

Deep learning is mainly based on accurate understanding, in-depth exploration, and reflection and driven by learners' intrinsic motivation, to critically and autonomously learn new knowledge and new ideas, solve practical problems, and cultivate students' deep learning and exploration capabilities [1]. This model has an important role in promoting the cultivation of students' abilities, the shaping of students' character, and the improvement of students' quality [2]. Therefore, in junior high school music education teaching, teachers should reasonably use the deep learning model to improve students' comprehensive music ability [3]. Deep learning is a field of machine learning which is inspired by a neural structure [4]. These networks extract the features automatically from the dataset and are capable of learning any nonlinear function. That is why neural networks are called as universal functional approximators [5].

Deep learning is a learning model that helps students develop advanced thinking based on comprehension learning, efficiently solve practical problems, and thus critically learn new knowledge and ideas and integrate them into the original knowledge structure [6, 7]. This learning model places great emphasis on critical understanding, guiding students to scientifically integrate information and further reconstruct the body of knowledge [8]. This learning model incorporates learning objectives that allow students not only to know and understand but also to achieve advanced cognition, to reach advanced levels of cognition, to apply analysis, synthesis, and evaluation, to apply their knowledge flexibly and scientifically, and to develop advanced higher-order thinking [9].

In the music classroom, teachers encourage and guide students to reproduce music, i.e., to recreate musical works [10]. To achieve this, teachers need to organize in-depth teaching and instruction in music teaching. Each piece of music contains the emotions of the composer and the background of the composition [9]. In the middle school music classroom, if teachers simply teach students the basic knowledge and skills of music, they will not be able to help students grasp the connotation of music comprehensively and accurately, and they will not be able to realize the second degree of creation of music works. Therefore, teachers should guide students to actively participate in music activities and to learn in-depth in classroom teaching, so that they can develop students' music appreciation ability, improve their music literacy, enable them to form good creative ability, appreciation ability, and expression ability in music, effectively cultivate students' aesthetic interest,

and enable them to form an optimistic and upward-looking attitude towards life and to love life and expression more. Students can develop an optimistic and upward attitude towards life, love life more, and learn to live so that students can develop physically and mentally healthily and happily.

The rest of the paper is organized as follows. In Section 2, materials and methods are discussed followed by experiments and results in Section 3, and the paper is concluded in Section 4.

2. Methodology

The traditional machine learning methods used for audio recognition are decision trees, support vector machines, logistic regression, etc. Although the above methods can perform most of the classification tasks, feature extraction is difficult and thus the classification performance is poor. In recent years, deep learning methods based on neural networks have attracted more and more attention. It mainly refers to modeling the human brain and constructing a neural network to simulate the functions of the human brain. The human brain is a neural network made up of multiple neutrons; Similarly, artificial neural networks (ANNs) composed of multiple perceptrons provide new solutions for audio recognition tasks, where the common forms of convolution for processing audio include one-dimensional convolution [12] and two-dimensional convolution [13–15]. One-dimensional convolution is commonly used for feature extraction of sequence data, and the data are often taken as input with one-dimensional audio time-domain signal data, which will be referred to as raw audio data later; two-dimensional convolution is commonly used for feature extraction of two-dimensional data, and the conversion spectrogram operation is often performed on raw audio data during audio data processing.

In this paper, two convolution-based deep learning models, Breath1d and Breath2d, are designed and constructed, and a multilayer perceptron (MLP) is used as a benchmark method for performance evaluation, and a long short-term memory (LSTM) network [15] is applied for the classification task. A multilayer perceptron (MLP) is a class of feedforward artificial neural network (ANN). An MLP consists of at least three layers of nodes: an input layer, a hidden layer, and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. Figure 1 shows the training and testing process of the deep learning model on the Breath dataset.

2.1. Introduction to General Deep Learning Models. A simple form of MLP is given in Figure 2, where the preprocessed data are fed into the network and trained in the implicit layer, and the prediction layer is the output of the softmax function in the output layer. In addition, an LSTM model is used to classify the Breath dataset. In this paper, a cell module with 50 hidden vectors is used as the LSTM layer, followed by a dense layer as the classification output. Based on the input characteristics of the Breath dataset, the input feature length of the LSTM network is 513 and the feature step length is 169.

2.2. Breath1d Neural Network Model. The structure of Breathh1d, a one-dimensional convolutional neural network designed in this paper, is shown in Figure 3. In Figure 3, there are 3 similar blocks and 1 fully connected layer in the hidden layer. Each block consists of a convolutional layer, a pooling layer, and a dropout layer. The parameters of the Breath1d model are constructed as shown in Table 1.

Since the 1D convolution method is to process the original audio file, 1D convolution sets two convolutional layers in the first block, each layer is 9 or 16 convolution kernels in length to facilitate the acquisition of audio features through a larger sensory field; the number of convolution kernels is increased to 64 in the second block, and the number of convolution kernels is increased to 256 in the third block, and the length of convolution kernels is reduced to 3 to extract high-dimensional features; the activation function of each layer is chosen as Relu. Relu is chosen for each layer of activation function, and the dropout method is added in each block to avoid overfitting.

2.3. Breath2d Neural Network Model. The one-dimensional convolutional neural network Breath1d designed in this paper is shown in Figure 3. Breath2d network structure is shown in Figure 4.

In Figure 4, the input layer is fed with transformed two-dimensional spectral data. The data are convolved in each layer and then batch normalized for faster convergence. The final layer uses a fully connected layer that outputs a probability value or returns a list as a one-hot by a softmax function. Parameters of the Breath2d network at each level are shown in Table 2.

The two-dimensional convolution requires the input data to be the frequency-domain features obtained by the short-time Fourier transform (STFT), where the STFT window is set to 1,024 and the sliding step is 256-second duration of each audio. The short-time Fourier transform is a Fourier-related transform used to determine the sinusoidal frequency and phase content of local sections of a signal as it changes over time. STFT is frequently used to analyze music. The dimensions of the first two convolution kernels of the Breath2d model structure are set to (10, 4) according to the input dimensions to maximize the ability of the sensory field to acquire data features; a maximum pooling layer is added after each convolution to reduce the number of subsequent computational parameters.

3. Experiments

In this paper, we perform experimental classification of the bamboo flute dataset to find the best classification method and classify the experimental categories as follows:

Experiment 1 dichotomized ventriloquism, trill, and tongue technique with a flat blow to explore the classification performance of similar techniques.

Experiment 2 performed dichotomous classification of triple and double spit to explore the effectiveness of neural network training for short interval period sounds.

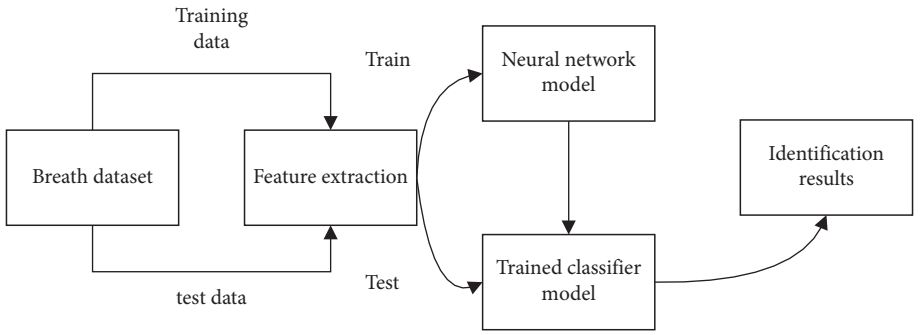

Figure 1: Training and testing process of deep learning model.

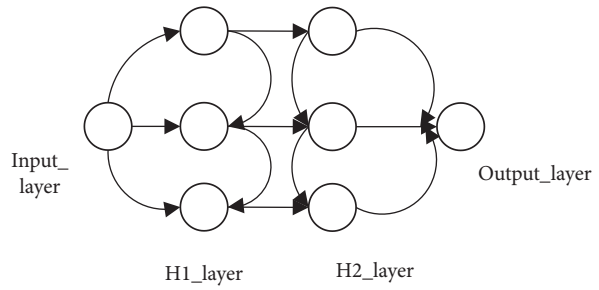

Figure 2: General deep learning model structure.

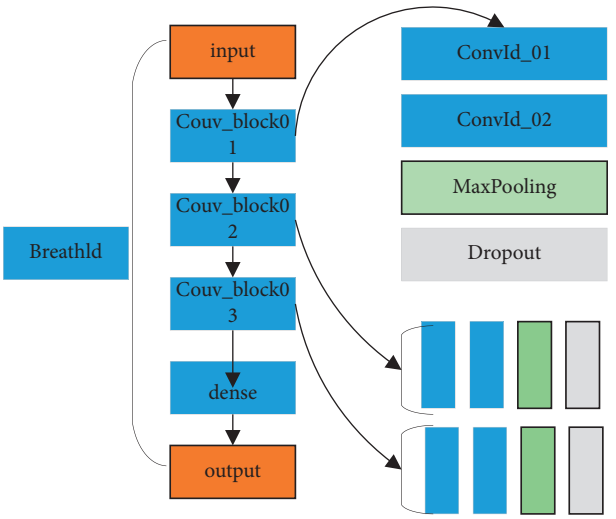

Figure 3: Network structure of Breath1d.

Table 1: Breath1d network parameters at each level.

Network layer	Number of convolution kernels	Convolution kernel/pooling parameter
Convld_01	16	9
Convld_02	16	9
Maxpooling_01	—	16
Convld_03	64	9
Convld_04	64	9
Maxpooling_02	—	4
Convld_05	256	3
Convld_06	256	3
Maxpooling_03	—	3

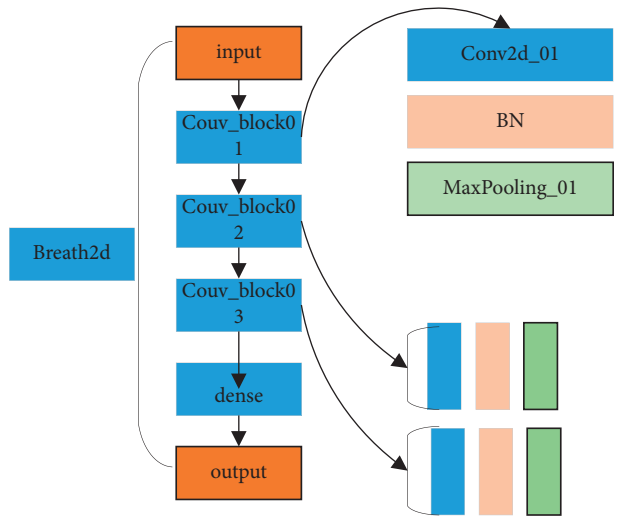

Figure 4: Breath2d network structure.

Table 2: Parameters of Breath2d network at each level.

Network layer	Number of convolution kernels	Convolution kernel/pooling parameter
Conv2d_01	16	(10, 4)
Maxpooling_01	—	(4, 2)
Convld_02	32	(10, 4)
Maxpooling_02	—	(4, 2)
Conv2d_03	64	(3, 3)
Maxpooling_03	—	(2, 2)
Conv2d_04	128	(3, 3)
Maxpooling_04	—	(4, 2)

Experiment 3: 6 techniques were distinguished: flat blow, ventral vibration, trill, flower tongue, double spit, and triple spit, and the best classification model was selected by comparison.

Experiments 1 and 2 are two classification tasks, and Experiment 3 is a six-classification task. The neural network parameters are constructed according to the parameters in Tables 1 and 2, where the first two layers of the two-dimensional convolutional network are filled with the same dimension; the activation layers of the convolutional network are activated by the Relu function, except for the fully connected layer, which is activated by the softmax activation function. The softmax function is used in the final layer of a neural network-based classifier. It is often used as the last activation function of a neural network to normalize the output of a network to a probability distribution over predicted output classes.

Experiment 1 used MLP as the benchmark method and applied the Breath1d model, Breath2d model, and LSTM model to train the Breath training set data. The classification results of Experiment 1 are shown in Table 3.

The comparison results of Experiment 1 are shown in Figure 5.

Comparing Figure 5, we can find that the classification accuracy of bamboo flute audio processed by Breath1d network model is significantly higher than that of the benchmark method on the test set but lower than that of Breath2d and LSTM models; whether the input of Breath2d model is processed by MFCC or not has little effect on the classification results of ventriloquism and flat blow, there is a big difference in the classification of tremolo and flat blowing and flat-tongue blowing. MFCC has a great influence on the classification of ventriloquism. In general, the Breath1d model cannot fully extract the required features for this dataset, and the LSTM performance is equal for the three categories, while the Breath2d + MFCC model has the highest accuracy in the test set, indicating that the MFCC features are effective for this audio classification task. The classification accuracy of Experiment 1 is shown in Table 3.

Experiment 2 is an experiment that explores short-interval periodic audio. The features of this type of audio are sensitive to the duration interval, and the features extracted by the convolutional neural network are often less dependent on the audio duration, so it is difficult to classify triple and double spit. Experimental tests show that the classification of convolutional neural networks is poor, while the classification of triple and double spit trained by the LSTM network model is better, with an average score of about 90% for the accuracy of 10 model tests. It can be seen that the task of vocalization classification can be properly solved by applying the LSTM network model.

In this paper, we continue to explore the classification ability of the neural network model for the full set of Breath

Table 3: Classification accuracy of Experiment 1.

Network	Abdominal shock	Trill	Flower tongue
MLP	0.64	0.66	0.64
LSTM	0.88	0.86	0.58
Breath1d	0.82	0.84	0.50
Breath2d	0.88	0.86	0.86
Breath2d + mpcc0.64	0.88	0.92	0.90

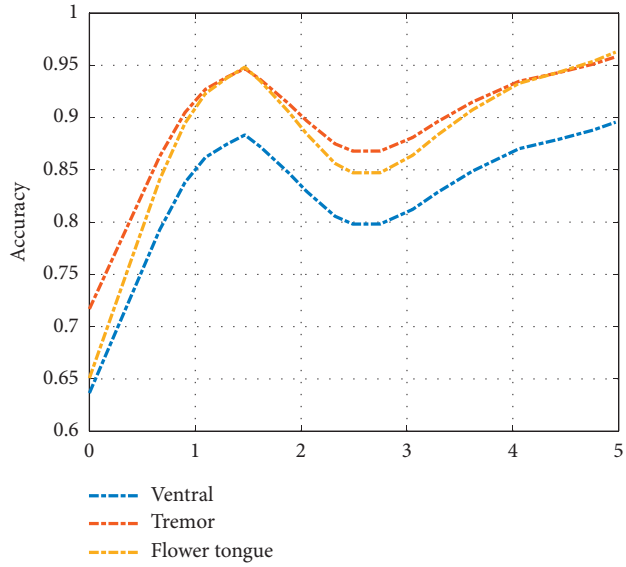

Figure 5: Result graph of Experiment 1.

in Experiment 3 based on the experience from Experiments 1 and 2. In addition to using the model from Experiment 1, Experiment 3 introduces the data enhancement method of pitch change and time stretching [15] and fuses Breath2d with Breath1d, i.e., the output of the Breath2d model is flattened and used as the input of the Breath1d model to form a new network. These operations effectively improve the predictive power of the model.

Since Experiment 3 was a multiclassification task performed on all the tricks within the dataset, the modified task was a six-classification one. During the experiment, it was found that although the neural network model got good convergence in the training phase, it did not perform well on the test set, indicating that the direct application of the data within the dataset for training on the existing amount of data did not yield the desired classification results. The reasons for this are as follows: (1) the amount of training data in the dataset is relatively small; (2) there is unavoidable noise in the recording, which produces artifacts that are not skillful audio features and interferes with the model's extraction of features; and (3) the timbre of the instruments in the test set differs from that in the training set, which depends on the quality of the instruments. To address this situation, this paper processes the data used based on the audio data enhancement method, which indirectly increases the diversity of the training data set by randomly scaling the spectrogram of the training data.

The results of Experiment 3 are shown in Figure 6.

From the experimental results in Figure 6, it can be seen that if the model is trained directly without introducing data enhancement methods, the ability of the model to correctly classify tricks is low, generally below 0.800. Although the classification accuracy can be improved to 0.807 by using the MFCC feature extraction method, the improvement effect is not obvious. To address the problem of insufficient data, this paper adds a data enhancement method of random time stretching and pitch adjustment to the best-performing Breath2d model. The performance of the model on the test set was also improved due to the increased data diversity. For predictions performed on the original test set, the average accuracy improved by 0.026. To explore the optimal capabilities of the neural network, this paper further improved the classification performance by connecting the Breath2d to the Breath1d model, and the addition of the data enhancement method to the Breath2d and Breath1d fusion models resulted in 0.040 improvements over the original fusion model, with the classification accuracy reaching the highest (0.913).

The simultaneous use of data enhancement methods and the fusion of Breath2d and Breath1d models can effectively improve the classification prediction accuracy. The reasons for this analysis are as follows: the introduction of the data augmentation method improves the accuracy of the model classification due to the characteristics of the Breath dataset. Although the training and test sets of the Breath dataset were

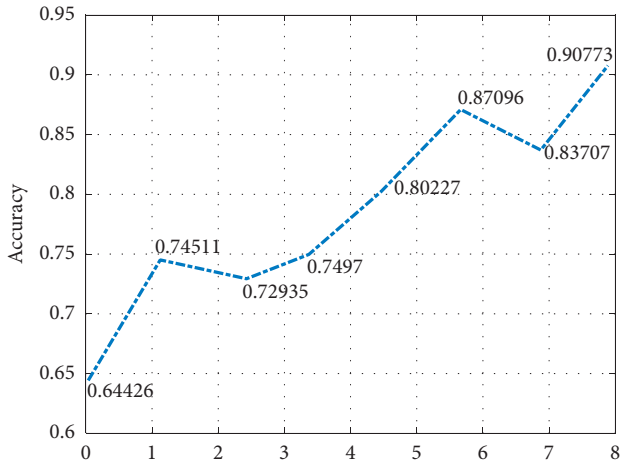

Figure 6: Comparison of the results of Experiment 3.

recorded separately, the main features were determined by the pitch distribution of the audio. The data enhancement method using pitch alteration with time stretching is essentially changing the pitch of the bamboo flute technique. The added portion of the data has similar features to the data in the test set, so the accuracy of the model classification increases accordingly when using the test set for prediction. Fusing the Breath1d and Breath2d models increases the complexity of the model, i.e., it increases the fit of the data in terms of similar feature prediction and also results in a corresponding increase in the prediction accuracy of the test set.

Experiment 3 also illustrates that this dataset is complete for the bamboo flute playing skill classification task and that the effect of different bamboo flute timbres on the classification of bamboo flute skills can be largely ignored with simple processing, thus completing the bamboo flute skill classification task accurately.

4. Conclusion

In this paper, a dataset dedicated to bamboo flute techniques is constructed and two neural network models are proposed. The LSTM network can extract features better for the classification of short-period interval audio. The analysis of the improved accuracy also verifies the completeness of the Breath dataset. This paper explores the automatic recognition of bamboo flute techniques, but due to the perception of the art form, there is still a large distance to truly achieving automatic appreciation of music. In future work, the technique classification can be extended to other musical instruments, and it can be combined with detection algorithms to complete the global technique analysis of a complete piece of music, and a source separation algorithm can be used to extract pure bamboo flute tones from existing music to expand the existing dataset and form a more diverse training dataset.

Data Availability

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

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

This study was supported by the Shanxi Provincial Education Reform Project (Reform of practical teaching system and evaluation mechanism of music major) (Project No. 2019GJ03), Shanxi Normal University.

References

[1] S. Scott, "A constructivist view of music education: perspectives for deep learning," *General Music Today*, vol. 19, no. 2, pp. 17–21, 2006.

[2] M. Świechowski, H. Park, J. Mańdziuk, and K.-J. Kim, "Recent advances in general game playing," *The Scientific World Journal*, vol. 2015, Article ID 986262, 22 pages, 2015.

[3] Z. Tan, J. Yuan, and H. Bao, "Estimation of crowd density based on deep convolutional neural networks," *International Journal of Engineering Intelligent Systems for Electrical Engineering & Communications*, vol. 24, no. 3-4, pp. 131–138, 2016.

[4] W. Hryniewska, P. Bombiński, P. Szatkowski, P. Tomaszewska, A. Przelaskowski, and P. Biecek, "Checklist for responsible deep learning modeling of medical images based on COVID-19 detection studies," *Pattern Recognition*, vol. 118, no. 2, Article ID 108035, 2021.

[5] G. P. Liu, J. J. Yan, Y. Q. Wang et al., "Deep learning-based syndrome diagnosis of chronic gastritis," *Computational and Mathematical Methods in Medicine*, vol. 2014, Article ID 938350, 8 pages, 2014.

[6] D. A. Whitenack and P. E. Swanson, "The transformative potential of boundary spanners: a narrative inquiry into preservice teacher education and professional development in an NCLB-impacted context," *Education Policy Analysis Archives*, vol. 21, p. 19, 2013.

[7] D. Sun, J. Wu, H. Huang, R. Wang, F. Liang, and H. Xinhua, "Prediction of short-time rainfall based on deep learning," *Mathematical Problems in Engineering*, vol. 2021, Article ID 6664413, 8 pages, 2021.

[8] C. N. Phyo, T. T. Zin, and P. Tin, "Deep learning for recognizing human activities using motions of skeletal joints,"

IEEE Transactions on Consumer Electronics, vol. 65, no. 2, 2019.

[9] T. Poddig and A. Unger, "On the robustness of risk-based asset allocations," *Financial Markets and Portfolio Management*, vol. 26, no. 3, pp. 369–401, 2012.

[10] Y. Zhang and U. Nauman, "Deep learning trends driven by temes: a philosophical perspective," *IEEE Access*, vol. 8, pp. 196587–196599, 2020.

[11] Z. You, J. Fang, and I.-T. Lu, "Out-of-band emission suppression techniques based on a generalized OFDM framework," *EURASIP Journal on Applied Signal Processing*, vol. 2014, no. 1, p. 74, 2014.

[12] Z. He, X. Zhu, and J. Li, "Modeling and simulation for the operation process of cold-chain logistics distribution center based on flexsim," in *Liss 2013*, pp. 277–282, Springer, Berlin, Germany, 2015.

[13] A. Christodoulou and J. Osborne, "The science classroom as a site of epistemic talk: a case study of a teacher's attempts to teach science based on argument," *Journal of Research in Science Teaching*, vol. 51, no. 10, pp. 1275–1300, 2014.

[14] Z. Jie, S. Pan, and D. Yan, "The effect of a simulation-based training on the performance of ACLS and trauma team of 5-year medical students," *Lecture Notes in Electrical Engineering*, vol. 269, pp. 253–263, 2014.

[15] S. N. Güngr, D. Z. Zer, and M. Zkan, "A study on the evaluation of scientific projects of primary school students based on scientific criteria," *Asia-Pacific Forum on Science Learning and Teaching*, vol. 14, no. 2, p. 40, 2013.