Research on Music Genre Classification Based on Residual Network

Zhongwei Xu[1], Yuan Feng[1(✉)], Shengyu Song[1], Yuanxiang Xu[1], Ruiying Wang[2], Lan Zhang[1], and Jiahao Liu[1]

[1] College of Information Science and Engineering, Ocean University of China, Qingdao 266005, China
fengyuan@ouc.edu.cn, {Songshengyu,liujiahao6266}@stu.ouc.edu.cn
[2] Teaching Center of Fundamental Courses, Ocean University of China, Qingdao 266005, China

Abstract. With the rapid development of information technology, the number of songs is exploding, so the classification of music genres is a very challenging task, and at this stage, the implementation of automated classification of music genres is also a relatively popular scientific research topic. Mobile devices are all over people's lives and have brought great convenience to people's life and work, making it possible to work anywhere and anytime. However, the special characteristics of mobile devices require high model requirements, which are difficult to be realized by traditional models. We hope to use deep learning to automatically identify and classify music, and use the Mobilenet model to achieve lightweight music classification on mobile and improve the classification accuracy. In this paper, we mainly use Free Music Archive dataset for experiments, based on resnet101 network model and MobileNet model for music genre classification, mainly use Short Time Fourier Transform (STFT) and Mel Frequency Cepstrum Coefficient (MFCC) for music feature extraction, improve the data pre-processing, and compare with other model methods were compared, and the accuracy rate was about 7% higher than the traditional CRNN method, and better results were achieved. On the implementation of the lightweight model for mobile, the size of the parameters of the model trained by MobileNet is only 4% of the best model in this paper, and has a high accuracy rate.

Keywords: Residual Network (ResNet) · MobileNet · Depth learning · Short Fourier Transform (STFT) · Mel Frequency Cepstral Coefficient (MFCC)

1 Introduction

In the age of information technology, audio is everywhere. The gym may be playing a powerful DJ; the bookstore is playing beautiful light music; the concert hall is playing a rhythmic concerto; and there is a lullaby before bed…… Different scenes apply different types of music, and different genres of music are loved by different people. Nowadays, due to the rapid development of science and technology, the production of music has become easier and there are a large number of music lovers who are engaged in music,

Y. Chenggang et al. (Eds.): MobiMedia 2022, LNICST 451, pp. 209–223, 2022.
https://doi.org/10.1007/978-3-031-23902-1_16

and with it, the number of music tracks has skyrocketed. There are various music playing software, and the classification of music genres is more refined, and even personalized song list is pushed according to individual listening habits, but it is a great challenge for backend workers to classify music genres manually. Combined with the popularity of mobile devices and the increasing speed of computing, many lightweight tasks can already be done on mobile. However, traditional neural network training parameters and computing volume are relatively large, which is difficult to implement in mobile.

In this paper, we mainly use different feature extraction methods and different models to classify music genres. Different music genres have their unique musical melodies, but some music has a certain degree of similarity and is still difficult to distinguish, for example, pop is interspersed with other types of rhythms, so it is also more difficult to achieve in this dataset classification. Some music has more obvious features, and the classification effect is better (for example, hip-hop instruments are easier to distinguish). At present, MFCC and STFT are more commonly used music feature extraction methods, and have been more widely used in music classification, and have also achieved better results.

We have studied the way of music processing based on MFCC and STFT feature extraction. The traditional method is to feature extract the whole music and then input to the model for training and learning, but some of the audio is long and there may be a lot of data loss if made into feature images for learning (the size of the image display is limited), and the size of the images that can be input to most models for training is also limited. There is no continuity between different audios in music genre classification, so the training effect is not very good. In this paper, the original audio is re-sliced into 4s segments and each segment is overlapped by 50%, and then each segment is subjected to feature extraction to reduce data loss and make the audio segments have some continuity and relevance.

In the use of training models, we selected popular classification method models (ResNet, GoogleNet, MobileNet, etc.) for the unsliced audio to select different feature extraction methods for training and learning, and initially judged which model corresponding to which feature could achieve relatively good results, and we found that STFT was able to have a good accuracy rate in each model training. And using the Resnet101 model and STFT feature extraction method, the best results in training can be achieved with 71.8% accuracy after training with sliced data. And we used MobileNet network to achieve the feasibility of music genre classification in mobile, using two data processing methods achieved better results accuracy reached 58.2% and 60.9%, and the final model parameter size is only about 4% of the traditional model parameters.

1.1 Related Work

In recent years, the classification of music genres is a relatively important topic, and more and more people are devoted to related research fields. Audio feature extraction methods mainly include STFT, MFCC, MEL, etc., whose feature extraction methods also play an important role in different research fields Tao et al. (2020) [1] and Liu et al. (2019) [2]. With the development of information technology, the classification methods of music genres have been continuously pushed from using traditional classical classifiers Basili et al. (2004) [3]; Kostrzewa et al. (2018) [4] and Silla, C.N (2008) [5] to the field of deep

learning Choi, K et al. (2017) [6]; Kereliuk et al. (2015) [7] and Oramas et al. (2017) [8], and the current ones are more popular are convolutional neural networks Kim et al. (2018) [9]; Lee et al. (2017) [10] and Lim et al. (2018) [11] and convolutional recurrent neural networks Choi et al. (2017) [12] and Gunawan et al. (2019) [13]. There are also combined methods of different neural networks to explore for music genre classification, such as CNN + LSTM and VGG16 + LSTM [14]. There are also many results on music genre classification using different audio feature extraction methods and combining different deep learning methods. Our previous explorations using models such as deep learning [15] and convolutional neural networks [16] found that the related models have some effect on music genre classification, and the sensitivity of different models to audio features is not the same.

Different datasets have different numbers of music categories, some with different lengths of music, and different criteria and conditions, and the trained results are not always reliable. In this paper, we use a standardized dataset with equal time for each song and compare the best possible effect with the experimental results of other works.

1.2 Contribution and Thesis Structure of the Paper

The main contribution of this paper is that firstly, we select different feature extraction methods for unsliced audio from popular classification models (Resnet, AlexNet, GoogleNet, MobileNet, etc.) and compare the experimental results with other methods to select the best model and feature extraction method. Secondly, the audio is fragmented and each adjacent segment is stacked by 50% on top of the best model and feature extraction method. Finally, the relevant parameters of the model are adjusted and a regularization optimization method is added to the model to prevent overfitting and an attention mechanism is added to improve the model training effect. In this paper, the MobileNet model is used to achieve a lightweight mobile music genre classification with good results.

The datasets and related settings used in this paper are described in Sect. 2; the main methods and depth models used are presented in Sect. 3; the experimental content and related analyses are presented in Sect. 4; and Sect. 5 concludes and outlooks.

2 Data

2.1 Data Sets and Hardware Settings

The Free Music Archive Dataset (FMA) has three main categories of varying sizes (small, medium, and large), and we choose the small dataset. The small subset of the FMA dataset contains 8000 tracks, each 30 s long, divided into 8 top categories each with 1000 audio is a balanced dataset, and the main categories are hip-hop, folk, experimental, international, instrumental, electronic, pop, and rock [17]. Each audio sample type is 44100 Hz dual-channel.

All experiments were conducted on a desktop computer with the following hardware configurations: CPU - Intel Core(TM) i7-9700K, RAM - 32GB RAM, GPU - NVIDIA GeForce GTX 2080 SUPER 8GB GDDR5.

2.2 Additional Settings

Data Normalization
As Eq. 1 shows the normalization method of data to channels:

$$O_{normalized} = (I - mean)/std \tag{1}$$

I is the input STFT spectrogram, mean is taken as [0.485, 0.456, 0.406], std is taken as [0.229, 0.224, 0.225] normalization can speed up further calculations to an order of magnitude and reduce the error rate [18].

Computer deep learning is indeed powerful in classification, but sometimes it is often prone to overfitting in order to achieve the expected or better results, especially when we set the relevant audio clips to overlap by 50%, so here we use the L2 regularization method when calling the resnet101 model, and we choose the Adam optimizer to set the parameter weight_decay = 0.001, i.e., the weight of the regularization is set to lambda = 0.001.

Data Pretreatment
The audio was read at a sampling frequency of 44100 Hz. The audio was segmented into audio clips of 4-s duration. To make each audio clip relevant and have continuity, we segmented the audio in such a way that each clip took the last 50% of the previous clip and accounted for the first 50% of the latter clip. This data processing both increases the amount of data and achieves strong correlation of the clip audio and enables more features to be learned, which facilitates computer recognition to improve learning efficiency.

Feature Extraction
We mainly extract two different features for each post-slice audio: Short Time Fourier Transform (STFT) and Mel-frequency spectral coefficients, and call different models to compare their accuracy, and then to determine which feature extraction and corresponding model are highly accurate.

3 Method

3.1 Short-Time Fourier Transform (STFT)

As shown in Fig. 1, the vein diagram of STFT image spectrum feature extraction is shown. Short-time Fourier transform is widely used in the field of music classification, and can better reflect the features of different music types. In the figure, the original audio is first binned (50% overlapping between adjacent segments), and the STFT features are extracted from the binned audio to generate the picture. We set n_fft: i.e., FFT window size to 1024 and hop_length: i.e., frame shift, to 512 here.

3.2 Mel-Frequency Spectrum Factor

A schematic diagram of the pulse of MFCC picture spectrum feature extraction is shown in Fig. 2. (MFCC) is a further extension of Mel spectrogram, Mel-Frequency cepstrum coefficients [19] is another representative method after audio clip spectrum compression. In the figure, the original audio is first binned (50% overlapping between adjacent clips), and the feature extraction of the binned audio is performed to generate the image.

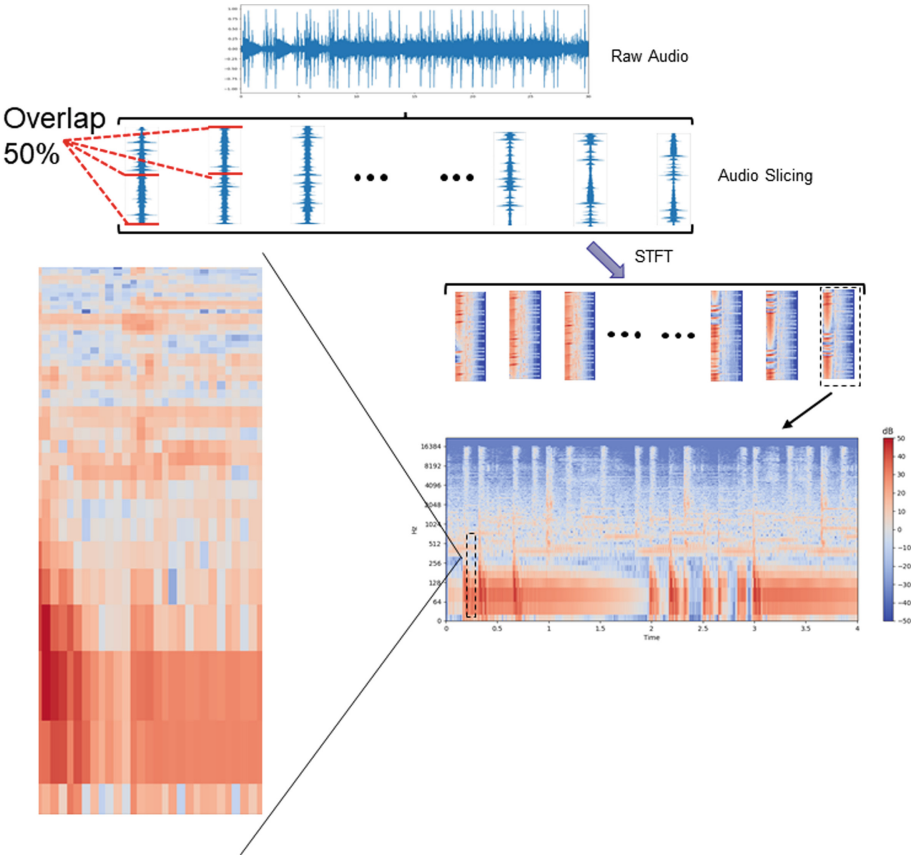

Fig. 1. STFT blue picture for 4 s segmentation middle 50% overlay

3.3 Deep Learning Models

As shown in Fig. 3, it is a network architecture of the ResNet 101 model. In the experiment, you can enter the MFCC or STFT feature to extract the generated picture, first pass through a convolution layer Stride 2, and then pass one 3 × 3 maximum sample layer Stride 2, followed by four residual blocks, and finally the average sample The layer and the full connecting layer are used as the output, as well as SoftMax processing.

The letters of the parameters accumulated below Layer1-Layer4 in Resnet101, we adjusted the number of layers of different Layers on the basis of the original parameters, respectively, and did the following sets of experiments, the audio feature used is STFT, we added and deleted the number of layers of different Layers respectively. As shown in the table below, the variation of the number of layers has a certain influence on the training accuracy, which does not mean that the more the number of layers is, the better the effect is, and the implementation of serial number 9 and 10 is better in comparison (Table 1).

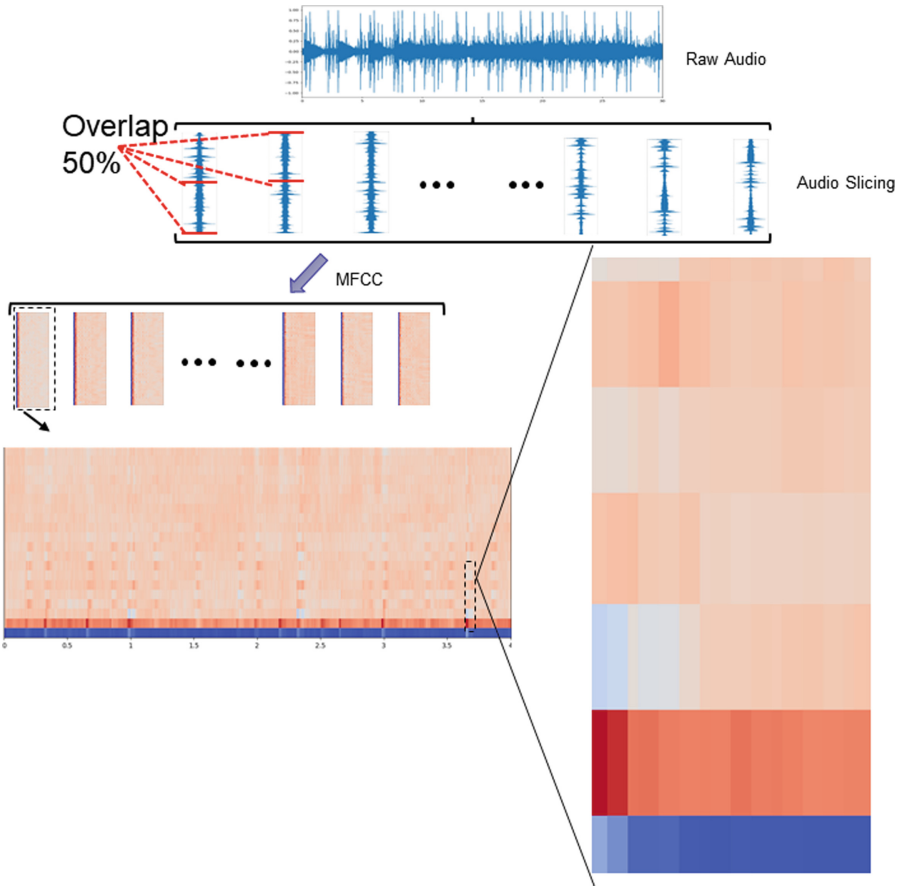

Fig. 2. MFCC blue picture for 4 s segmentation middle 50% overlay

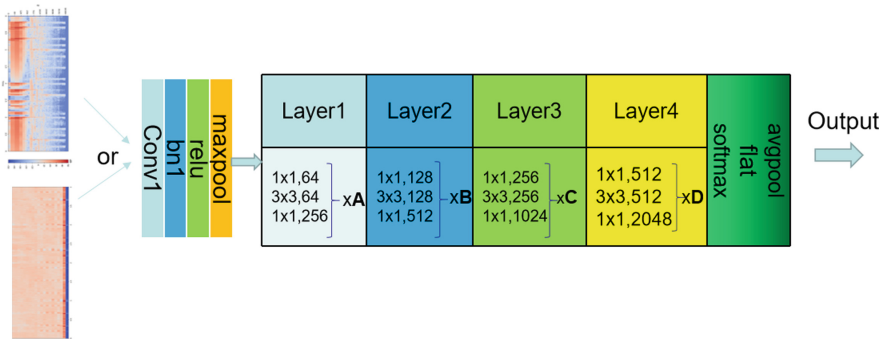

Fig. 3. ResNet

Table 1. Experimental results of Resnet101 parameter adjustment

No	A	B	C	D	Layers	Accuracy (%)
1	3	8	23	3	113	62.8
2	6	4	23	3	110	61.5
3	3	4	32	3	128	61.7
4	4	6	25	6	125	63.4
5	5	6	26	5	128	62.1
6	4	5	24	4	113	61.7
7	4	6	25	6	125	57.3
8	3	4	20	3	92	64.2
9	2	3	22	2	89	68.4
10	3	4	23	3	101	71.8

The following figure shows the comparison of the training accuracy with 10 parameter settings, the overall accuracy does not change much with the increase of the number of layers, and the accuracy is better with 89 and 101 layers.

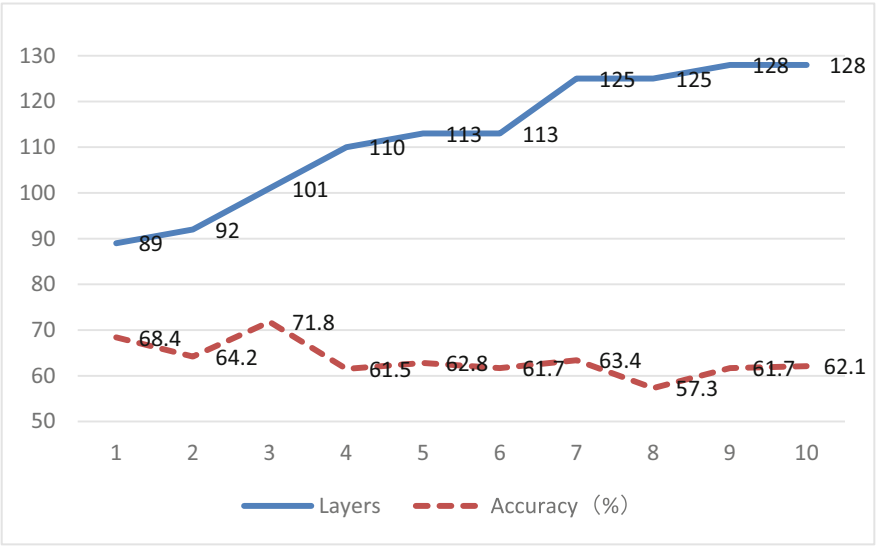

Fig. 4.

The MobileNet model is designed to achieve lightweight and high accuracy in mobile devices, which brings great convenience to life and work nowadays. Compared with the traditional convolutional neural network, MobileNet network has greatly reduced the model parameters and the amount of operations, but only a small decrease in accuracy.

As shown in Fig. 5, the size of MobileNet model is only 16.2MB, which is about 4% of the size of resnet101 model, and the training accuracy of MobileNet model is good compared with other models as shown in Table 6, which is better to achieve light weight.

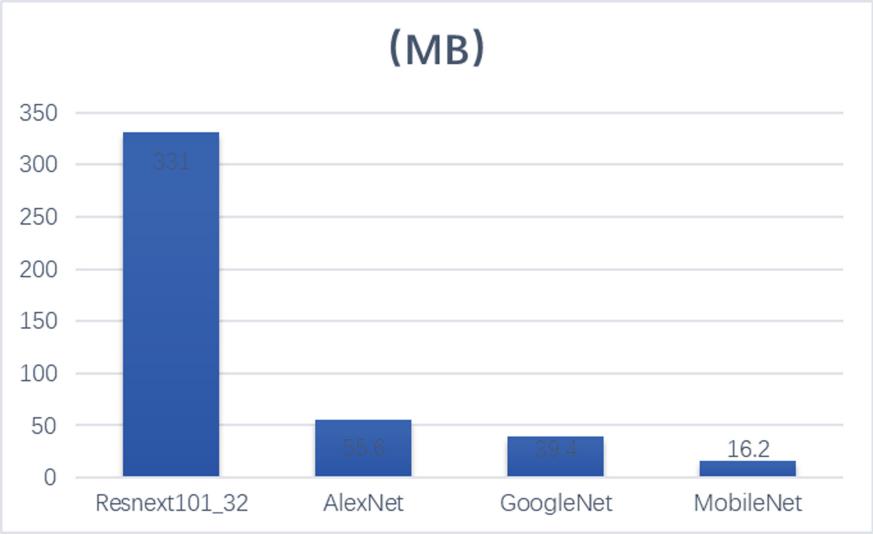

Fig. 5. Best results model parameter size comparison

4 Experiments and Analysis

As shown in Fig. 4 as the flow structure of this experiment, it is mainly divided into two processes, the first is the dataset unsliced processing, direct feature extraction of the original audio data, and then random division of the data, and then call the model for training and learning, save the best training effect in each model, and finally is the prediction and results; the other is the music first slice before the related processing (Fig. 6).

4.1 Quantification Results

The data is divided into three parts: training data, validation data and test data, which are divided into 80% training data, 10% validation data and 10% test data. The data pieces can be called randomly for training.

Before the images were binned, we performed feature extraction and processed to generate STFT images, MFCC images, chromaticity images, and power spectrograms, as shown in Table 2 The accuracy rate trained by extracting STFT features was high, so we trained the images extracted by STFT features using Resnext101_32x8d and MobileNet models, respectively, as shown in the table The accuracy rate was between 63.5% and 58.2%, and the F1 scores are 57.9% and 63%, respectively.

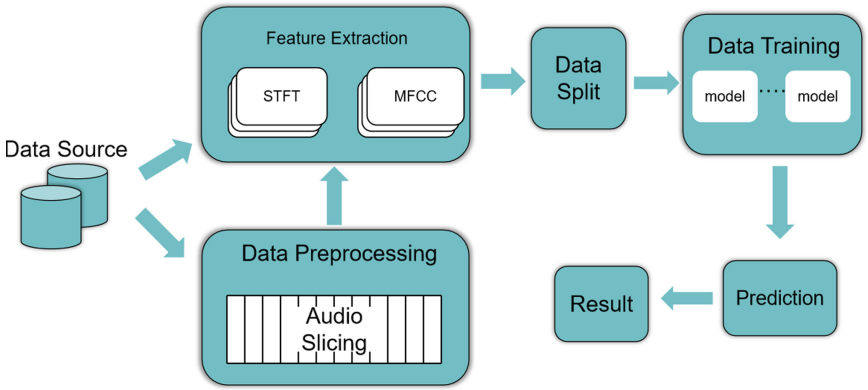

Fig. 6. Experimental flow chart

Table 2. Fma_small unsliced STFT feature learning

Model	Accuracy [%]	Precision [%]	Recall [%]	F1 Score [%]
Resnext101_32x8d	63.5	62.5	63.5	63
MobileNet	58.2	57.6	58.2	57.9

The audio after slicing and editing was also made into these three types of images, and the images extracted by STFT features were trained using Resnext101_32x8d and MobileNet models with accuracy at 71.8% and 60.9% with F1 scores of 71.8% and 60.8%, respectively. The accuracy of the classification effect after slicing the data was improved better (Table 3).

Table 3. Fma_small STFT feature learning after binning

Model	Accuracy [%]	Precision [%]	Recall [%]	F1 Score [%]
MobileNet	60.9	60.8	60.8	60.8
Resnext101_32x8d	71.8	71.6	72	71.8

Here you can see that the best model has an accuracy of 71.8% indeed has better results.

From here, we can see that there is a big gap between the two experimental results. Firstly, on the same feature extraction method, there is a big difference in the amount of data used by the two experiments. The total data volume without slicing is only 8000, the single category data volume is 1000, and the training set for each category is 900 validation set is 100; while the total data volume after slicing is 84183, the single category data volume is 11700, and the training set for each category is 10530 validation set is 1170, the experimental data volume difference is more than 10 times.

Secondly, in terms of data continuity and correlation, binning according to 50% of the front-to-back overlay makes a large number of data sets with large correlation, while unbinned audio each audio independently has little correlation.

Finally there is a large gap in the amount of data features. The model has a limit on the size of the input training images, and we process the audio according to the same image size, which inevitably results in the unsliced audio features being compressed, leading to a large loss of relevant data features. The slice overlay processing, which cuts the long audio into short audio, can show more information in the same size image during feature extraction, maximizing the use of data, and the overlay processing can eliminate the loss of data edge features, which can effectively improve the training efficiency.

4.2 Qualitative Results

As shown in Fig. 7 for the confusion matrix best model classification effect, the number of predicted correct results if they are all concentrated on the diagonal line, indicating the higher the accuracy, which is the best effect shown in the experiments, there is still a large difference in the music classification effect of different genre categories. The amount of validation data for each category is about 1170, and the number of validation sets for each category accounts for 20% of the total data for a single category, here

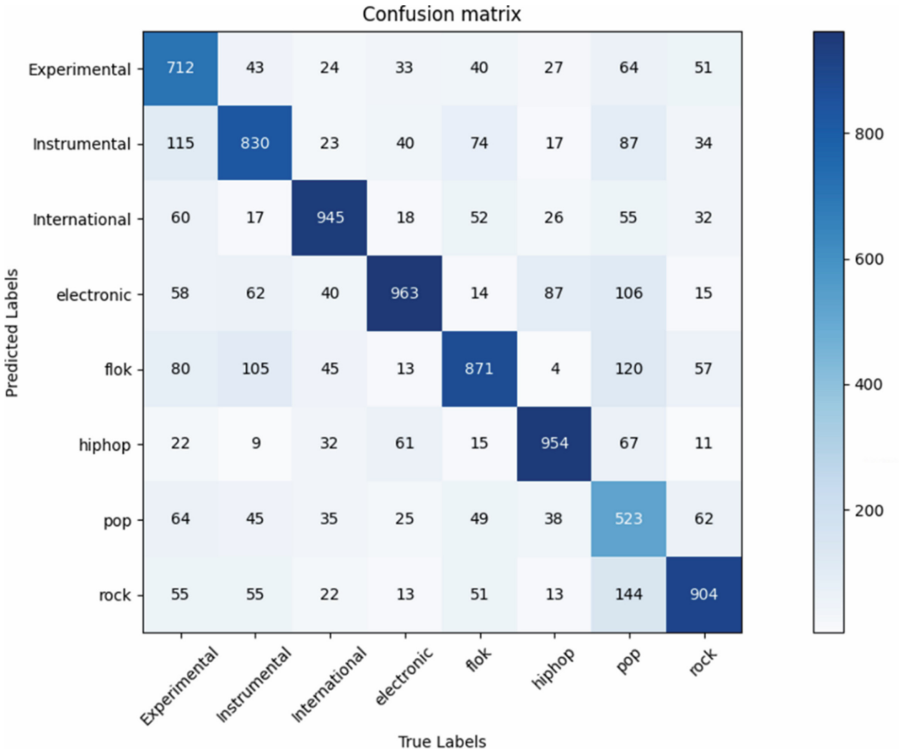

Fig. 7. Confusion matrix best model effect

we call the best training model for the validation set to generate the confusion matrix for testing, as shown in the diagonal line with the darkest color, the number of each category being predicted correctly is concentrated in the squares on the diagonal line, International, electronic and International, electronic and hip-hop are the best in terms of accuracy, pop is relatively poor because pop is more similar to other music, but overall the prediction accuracy is better.

As shown in Table 4 are the training effects of the best model in each genre category. Among them, hip-hop, international, electronic and rock achieved more than 80% accuracy, folk, experimental and instrumental also achieved between 60% and 80%, the worst result is pop only 44.7%, because pop is more special and complex, an audio contains features of almost other genres of music, it is the most difficult to classify, but compared to the paper [17] around 20% we have a large improvement.

Table 4. Best model for each genre of music training effect

	Exp	Ins	Int	Ele	Flok	Hiphop	Pop	Rock
Acc	60.9%	70.9%	80.8%	82.3%	74.4%	81.5%	44.7%	77.3%

As shown in Table 5 for the best model in each genre of music training effect, hiphop accuracy is the best at 81.5%, and all other genres except POP accuracy is above 65%.

Table 5. Best model for each genre of music training effect

Precision		Recall	Specificity
Experimental	0.716	0.611	0.965
Instrumental	0.68	0.712	0.952
International	0.784	0.81	0.968
Electronic	0.716	0.826	0.953
Flok	0.673	0.747	0.948
Hiphop	0.815	0.818	0.973
Pop	0.622	0.449	0.961
Rock	0.719	0.775	0.957

4.3 Comparison of Results

The table below shows a comparison with the relevant models in other Wang et al. (2019) [20]; Yi et al. (2019) [21]; Zhang et al. (2019) [22] and Kostrzewa et al. (2021) [17] papers and shows the relevant results and the best results of this study in black bolded part at the bottom, compared with other models the models and methods used in this paper achieve better results and have a better overall improvement.

Table 6. Compare the different model learning results of Fma_small with their own results. All values are expressed in %.

No	Model	Accuracy	Recall	F1-Score	Remarks
1	K-Nearest Neighbors [22]	36.4	–	–	
2	Logistic regression [22]	42.3	–	–	
3	Multilayer perceptron [22]	44.9	–	–	
4	Support vector machine [22]	46.4	–	–	
5	Original spectrogram [21]	49.4	–	–	
6	Harmonic spectrogram [21]	43.4	–	–	
7	Percussive spectrogram [21]	50.9	–	–	
8	Modulation spectrogram [21]	55.6	–	–	
9	FCN [20]	63.9	43	40.3	
10	TimbreCNN [20]	61.7	36.4	35	
11	End-to-end [20]	61.4	38.4	34.5	
12	CRNN [20]	63.4	40.7	40.2	
13	CRNN-TF [20]	64.7	43.5	42.3	
14	Ensemble 1 – vote [17]	56.4	54.8	54.9	
15	**Resnext101_32x8d**	**63.2**	–	–	**Power spectrum unsecured**
16	**Resnext101_32x8d**	**50.4**	–	–	**Chromaticity map is not separated**
17	**Resnext101_32x8d**	**47.9**	–	–	**MFCC unsliced**
18	**AlexNet**	**61.7**	–	–	**STFT unsliced**
19	**Resnext101_32x8d**	**63.5**	**63.5**	**63**	**STFT unsliced**
20	**MobileNet**	**58.2**	**58.2**	**57.9**	**STFT unsliced**
21	**MobileNet**	**60.9**	**60.8**	**60.8**	**STFT Sliced**
22	**Resnext101_32x8d**	**71.8**	**72**	**71.8**	**STFT Sliced**

The analysis of Table 6 allows us to conclude that, in general, the values of the results of this paper are similar to those of other studies. Rows 1–4 (Table 6) show the values obtained by the classical classification method [22], which are 25–35% lower compared to the values obtained by our best results. Rows 5–8 show the results for

CNN provided by different spectral maps, and rows 9–13 show the quantitative results of full 64 D. Kostrzewa et al. for convolutional neural networks, timbre CNNs, end-to-end music auto-labeling methods, CRNNs, and CRNNs with temporal and frequency dimensions [20], with 7–10% lower accuracy compared to the best results in this paper and with recall and F1 scores are much lower than the current study reaching about 29%. Summarizing all the results, it can be seen that the results achieved in this paper study are 7% higher than the state-of-the-art solution, and better scores are achieved for other parameters. The accuracy of the MobileNet model is also among the better model accuracy as shown in Fig. The effect of the data slicing and stacking process is improved and the model parameter size is smaller.

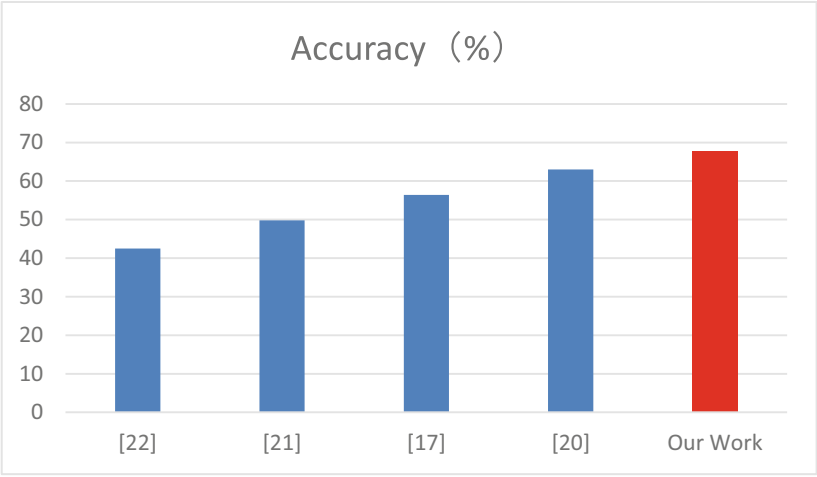

Fig. 8. Comparison of the average accuracy of the best models in the citation and this paper

As shown in Fig. 8, the average accuracy of the citations Kostrzewa et al. (2021) [17]; Wang et al. (2019) [20]; Yi et al. (2019) [21] and Zhang et al. (2019) [22] and the best model in this paper are compared, the blue part is the average of the citation model effect accuracy, and the red part is the average of the best model effect accuracy in our work, and the model in this paper is best in comparison.

5 Conclusion

We used different data processing methods for slicing and overlaying the data to increase the relevance and continuity of the data, and experimentally compared the selection of appropriate classification models and feature extraction methods. The quantitative and qualitative studies show that the obtained experimental results have a large improvement compared to the state-of-the-art methods. In addition, the best feature extraction method (STFT) and the best model (Resnet101) were screened in this study, and the Resnet model was studied and improved, and the classification quality was also improved better by using the STFT feature extraction method. In this paper, we also use MobileNet

model to classify music genres for mobile, and it has good accuracy compared with the best model.

The approach shown in this paper has the advantages of better data processing and feature extraction methods, a relatively new and advanced classification model, and the ability to automate the classification application with good classification results.

Future work can be focused on other deep learning model optimization improvements and other datasets for configuration experiments. The next step will be to determine the classification of the problems that occur in song singing.

Funding. This work was supported in part by the National Research and Development Program of China under 2020YFB1710401, and in part by the National Natural Science Foundation of China under Grant 61902367 and Grant 41976185.

References

1. Tao, H., et al.: An unsupervised fault diagnosis method for rolling bearing using STFT and generative neural networks. J. Franklin Inst. **357**(11) (2020)
2. Liu, C., et al.: Bottom-up broadcast neural network for music genre classification (2019)
3. Basili, R., Serafifini, A., Stellato, A.: Classifification of musical genre: a machine learning approach. In: ISMIR (2004)
4. Kostrzewa, D., Brzeski, R., Kubanski, M.: The classifification of music by the genre using the KNN classififier. In: Kozielski, S., Mrozek, D., Kasprowski, P., Malysiak-Mrozek, B., Kostrzewa, D. (eds.) BDAS 2018. CCIS, vol. 928, pp. 233– 242. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99987-618
5. Silla, C.N., Koerich, A.L., Kaestner, C.A.: A machine learning approach to automatic music genre classifification. J. Braz. Comput. Soc. **14**(3), 7–18 (2008)
6. Choi, K., Fazekas, G., Sandler, M., Cho, K.: Transfer learning for music classifification and regression tasks. arXiv preprint arXiv:1703.09179 (2017)
7. Kereliuk, C., Sturm, B.L., Larsen, J.: Deep learning and music adversaries. IEEE Trans. Multimedia **17**(11), 2059–2071 (2015)
8. Oramas, S., Nieto, O., Barbieri, F., Serra, X.: Multi-label music genre classifification from audio, text, and images using deep features. arXiv preprint arXiv:1707.04916 (2017)
9. Kim, T., Lee, J., Nam, J.: Sample-level CNN architectures for music auto-tagging using raw waveforms. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 366–370. IEEE (2018)
10. Lee, J., Nam, J.: Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging. IEEE Signal Process. Lett. **24**(8), 1208–1212 (2017)
11. Lim, M., et al.: Convolutional neural network based audio event classifification. KSII Trans. Internet Inf. Syst. **12**(6), 2748–2760 (2018)
12. Choi, K., Fazekas, G., Sandler, M., Cho, K.: Convolutional recurrent neural networks for music classifification. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2392–2396. IEEE (2017)
13. Gunawan, A.A., et al.: Music recommender system based on genre using convolutional recurrent neural networks. Procedia Comput. Sci. **157**, 99–109 (2019)
14. Ahmad, F., et al.: Music genre classification using spectral analysis techniques with hybrid convolution-recurrent neural network. Int. J. Innov. Technol. Explor. Eng. (IJITEE) **9**(1) (2019)

15. Li, C., Feng, Y., Sun, T., Zhang, X.: Long term Indian Ocean Dipole (IOD) index prediction used deep learning by convLSTM. Remote Sens. **14**, 523 (2022)
16. Sun, T., Feng, Y., Li, C., Zhang, X.: High precision sea surface temperature prediction of long period and large area in the Indian ocean based on the temporal convolutional network and internet of things. Sensors **22**, 1636 (2022)
17. Kostrzewa, D., Kaminski, P., Brzeski, R.: Music genre classification: looking for the perfect network. In: Paszynski, M., Kranzlmüller, D., Krzhizhanovskaya, V.V., Dongarra, J.J., Sloot, P.M.A. (eds.) ICCS 2021. LNCS, vol. 12742, pp. 55–67. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-77961-0_6
18. Sola, J., Sevilla, J.: Importance of input data normalization for the application of neural networks to complex industrial problems. IEEE Trans. Nucl. Sci. **44**(3), 1464–1468 (1997)
19. Mel Frequency Ceptral Coefficient (MFCC) tutorial. Available at: http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/
20. Wang, Z., Muknahallipatna, S., Fan, M., Okray, A., Lan, C.: Music classifification using an improved CRNN with multi-directional spatial dependencies in both time and frequency dimensions. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2019)
21. Yi, Y., Chen, K.Y., Gu, H.Y.: Mixture of CNN experts from multiple acoustic feature domain for music genre classifification. In: 2019 Asia-Pacifific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 1250–1255. IEEE (2019)
22. Zhang, C., Zhang, Y., Chen, C.: SongNet: Real-Time Music Classifification. Stanford University Press, Palo Alto (2019)