Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/132/2024.20632

Music Genre Classification: A Comprehensive Study on Feature Fusion with CNN and MLP Architectures

Xiang Li[1,5,*,†], Fan Li[2,6,†], Zhi Peng Lu[3,7,†], Ziyue Yang[4,8,†]

[1]Guangdong University of Finance and Economics, China
[2]Jiangnan University, China
[3]South China Normal University, China
[4]Brown University, Shanghai, 02912, China

[5]3334332647@qq.com
[6]1033210324@stu.jiangnan.edu.cn
[7]20213802009@m.scnu.edu.cn
[8]Ziyue_yang@brown.edu
*corresponding author
[†]co-first authors

Abstract. The classification of music genres has been a key focus in the field of music information retrieval (MIR), many researchers are also looking for a number of ways to solve this problem. Early approach Feature-based methods such as Support Vector Machines (SVMs) and K-nearest Neighbors (k-NN) are mainly used for handcrafted features such as Mel frequency cepstrum coefficients (MFCC). As convolutional neural networks (CNNs) are increasingly used to capture complex time-frequency patterns in audio data, the performance and accuracy of music genre classification has been dramatically improved.Recent research advances have been made in integrating attention mechanisms and mixing mechanisms, which further improves the accuracy of the genre classification model. But that's it. Methods often rely on a single type of input data, such as spectrograms or raw audio data, and are therefore limited. They are able to fully capture the multifaceted nature of music.In this research, we propose a novel multi-input neural network architecture that uniquely integrates a CNN for Mel spectrogram processing with a Multilayer Perceptron (MLP) for handling additional feature data extracted from CSV files. This two-branch approach can effectively help us understand music data more comprehensively while processing spectral and feature-based information. The GTZAN dataset was a great help for our experiments, and the details of this part will be explained later. In addition, our model effectively combines the advantages of both CNN and MLP methods, thereby improving the classification accuracy of multiple music genres. And finally, our results show that the system achieves [0.77] accuracy, support [10], recall [0.67], and F1 score [0.65].

Keywords: Music Genre Classification, Audio Features, CNN, MLP, GTZAN Dataset.

1. Introduction

Although the popularity of music platforms has allowed us to have more and more access to a large amount of music, it is very challenging to manage this huge amount of data. The first is the huge amount

DOI: 10.54254/2755-2721/132/2024.20632

of data on the music platform itself, and the second is the similarities between certain music genres make manual management impractical, inefficient, and error-prone. But we don't have to be so pessimistic, the field of music recommendation research is looking for better solutions to help us solve this thorny problem.

An efficient and objective classification method was developed to manage this data. We can categorize music by genre and delivering it to relevant users, which can significantly enhance the user experience and improve platform retention. Convolutional neural networks (CNNs) are often used in image classification tasks. [1,2] In this study, we used a similar approach to classify music genres. Specifically, we trained a model that can tag and boost based on the musical genre of the corresponding track efficiency in the classification, organization and management of music. Specifically, we start by using CNNs for model training, three different features (Mel spectrogram, MFCC, and chroma feature) are extracted as input to the model parameters.The MLP model was then also used to input the feature data into a CSV file for training. Finally, by connecting the outputs of the CNN and MLP branches for fusion, our model effectively combines the best of both worlds.

As a result, the classification accuracy of multiple music genres has been improved.

2. Related work

In 2002, Tzanetakis and Cook published a paper titled “Classification of Musical Genres from Audio Signals”, which first proposed a method for music genre classification that combined audio signal processing with traditional machine learning methods. They also developed the GTZAN dataset (which is also the dataset we use this time), which is still widely used today. They also applied classification algorithms such as KNN and GMM, input features such as MFCC for model training, and finally achieved an accuracy of 61% on GTZAN. Although the accuracy seems not high, it is undeniable that their work is still considered the pioneer and foundation of music genre classification [3-5]. In 2014, Dieleman and Schrauwen published the paper “Primitive Learning of Musical Audio”. This paper mentioned that they used convolutional neural networks to perform end-to-end learning on raw audio data, thus avoiding the complex and inefficient method of manually extracting features [6]. In 2017, Keunwoo Choi and his team published the paper “Convolutional Recurrent Neural Networks for Music Classification”. In this paper, they proposed a hybrid model combining CNN and RNN to complete the audio classification task. The hybrid model takes advantage of the feature extraction ability of CNN and the temporal modeling ability of RNN to further improve the classification performance. Finally, a higher accuracy rate (about 80%) was achieved on the GTZAN dataset [7]. In the same year, Salamon and Bello also published their paper. Although their research mainly focused on environmental sound classification, their advanced methods have also been widely used to deal with music genre classification tasks, especially in the field of data augmentation.This study shows that data augmentation is important in audio classification, and also shows us how expanding datasets can enhance the generalization ability of deep learning models, which has a significant impact on subsequent music genre classification research [8].In 2019, Miron et al. showed in "A Recurrent Neural Network Method for Music Genre Classification Using Audio Features" that the RNN algorithm has advantages in capturing changes in music rhythm and melody. They mainly used the method of the RNN algorithm to capture time series information in music. It has also achieved good results[9]. In 2022, Fontes et al. used the self-attention mechanism of Transformer to process long sequence audio data in "Classifying Music Genres Using Deep Learning Techniques" and achieved an accuracy of nearly 90%.

3. Background research

The research on music genre classification has been driven by deep learning. Existing research can automatically extract audio features and improve classification accuracy and efficiency by applying algorithms such as CNN and RNN [4]. However, we are currently facing some new challenges.First, existing datasets are usually small and difficult to capture the diversity of various types of music. However, as more and more music data emerges, we can also get richer data sets, which can help us solve the problem of insufficient data from the source. However, the ambiguity of music styles and the

DOI: 10.54254/2755-2721/132/2024.20632

confusion of genres are also major challenges.Faced with this challenge, researchers are exploring more flexible multi-label classification models. They try to use unsupervised and self-supervised learning methods to better handle the problems of unlabeled data and genre mixtures. In addition, recommendation systems are another field closely related to music genre classification. Although most current recommendation systems rely on music genre classification, they often fail to meet personalized needs. Therefore, future recommendation algorithms may focus more on combining user behavior data with classification results, and use collaborative filtering, graph embedding and other technologies to provide music recommendations that are more in line with personal preferences. But even if these technologies advance rapidly, existing classification methods still have limitations. Because although deep learning can automatically extract features, it is difficult to capture all the complexities in music. In addition, we lack data on niche music, and this lack of data will affect the classification of some niche music. There are also significant differences in music styles across cultures and regions, which also poses certain challenges to the global classification model. Lastly, real-time classification and recommendation in large-scale music libraries continue to be constrained by computational resources and time efficiency.

In summary, while research in music genre classification has made substantial progress, it still faces multiple challenges across data, technology, and application domains. Future research will need to continue advancing deep learning techniques and integrate big data and personalized information to improve classification accuracy and user experience further.

The main objective of this study is to come up with and implement a hybrid neural network model that effectively combines CNNs and MLPs to improve music genre classification accuracy. As previous research has demonstrated, traditional methods like Support Vector Machines and K-Nearest Neighbors, provide simplicity and interpretability but struggle to surpass 85% accuracy on the GTZAN dataset [10-11]. Although these methods serve as a baseline, their performance does not meet modern application demands. Recent advances, like the 1D residual CNN architecture, have improved genre classification but still face accuracy limitations, achieving only around 80% [12]. Other deep learning approaches, such as BP Neural Networks, have shown higher accuracy, outperforming traditional machine learning models by 3.9%–7.2%, yet face challenges with imbalanced data [13].

Our model seeks to overcome these limitations by introducing a multi-input system that utilizes Mel spectrograms in the CNN branch and feature data in the MLP branch. By leveraging the strengths of both architectures, we aim to enhance classification performance across multiple music genres. This research also aims to explore how deep learning models, specifically 2D CNNs combined with dropout layers [14], can be further improved by fusing the spectral and global features of music data. Additionally, incorporating foremost methods such as attention mechanisms and Transformers, which have shown robustness across multiple datasets [15], may be explored in future extensions of this work.

4. Methodology

4.1. Proposed Architecture

The proposed model is a multi-input neural network architecture that integrates a CNN with a Multilayer Perceptron (MLP), each tailored to handle different types of input data. Specifically, the CNN branch is designed to process Mel spectrogram data, while the MLP branch handles feature data extracted from CSV files. The features extracted by these two branches are then combined through a concatenation layer, with the final classification performed by a fully connected layer.

CNN Branch: The CNN branch processes the Mel spectrogram data, which has a shape of (128, 1300, 1)—128 Mel bands, 1300 time steps per sample, and 1 channel (grayscale image). This branch consists of three convolutional layers (Conv2D) with 16, 32, and 64 filters, each with a filter size of (3, 3). Each convolutional layer has a max-pooling layer (MaxPooling2D) for downsampling, with kernel sizes of (2, 6), (2, 6), and (2, 4). After the convolution and pooling, then a Flatten layer is adopted to convert the feature maps into a one-dimensional array, making them proper for input to the fully connected layer. A BatchNormalization layer is applied to normalize the input of the neural network, speeding up the

DOI: 10.54254/2755-2721/132/2024.20632

training process. The Dense part contains 128 neurons with a ReLU and L2 regularization (lambda=0.02). In the end, we adopted a Dropout layer to reduce overfitting, with 40% of neurons randomly deactivated.

- **MLP Branch:** The MLP branch processes the feature data extracted from the CSV file, with the input shape being (X_train_features.shape[1],), corresponding to the volume of features. The Dense in this branch are configured as follows: the first layer has 128 neurons with a ReLU activation function and BatchNormalization; the second layer has 64 neurons with ReLU activation and Dropout; the third layer has 32 neurons with ReLU and Dropout.
- **Feature Fusion and Output Layer:** The outputs from the CNN and MLP branches are concatenated (concatenation), merging the feature vectors into a larger, unified feature vector. This is followed by a fully connected output layer. The final layer is a fully connected layer with softmax activation, designed to predict one of the 10 music genres.

4.2. Data set

We adopted the GTZAN dataset:

In this dataset, every kind of music has 100 music files. There is a total of 1000 music files, meaning in a total of 10 kinds. The data set is highly standardized, all 16bit, 22050hz. [3]

Our system is designed to classify music from different sorts, and then accurately recommend this music to users who enjoy a particular genre. Specifically, when a piece of music is input into the system in the supported format, it can automatically extract features from the music and analyze its genre using a pre-trained model. The music is then organized and stored in the corresponding category, after which it is recommended to users who prefer similar tags. This reduces the time and effort users spend searching for music.

5. Design

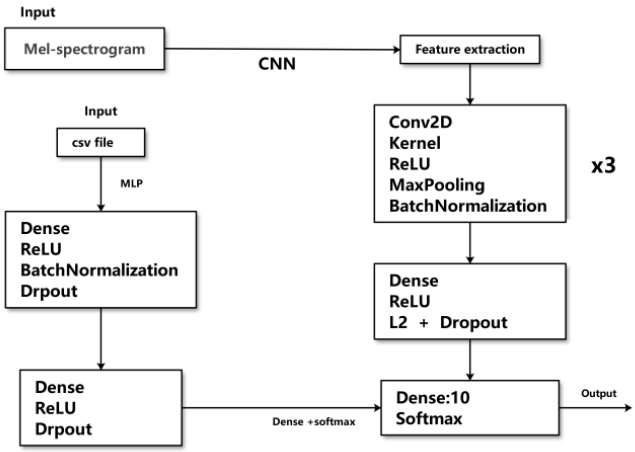

Figure 1. The image of designation

5.1. Data Preprocessing and Feature Extraction

In this study, the first step involved preprocessing and feature extraction from the audio data of various music genres. Using the librosa library, things like Mel spectrograms, MFCCs, and chroma features were extracted from the audio files. The code processes 30-second audio clips to generate spectral features. These features are normalized to ensure that the values of each spectrogram range between 0 and 1. To maintain consistency in the model input, all feature matrices are resized to a uniform length of 1300. During the preprocessing phase, the dataset needs to be divided into 3 parts, training, validation, and a test set, and the extracted features are saved as numpy arrays for use by the model.

DOI: 10.54254/2755-2721/132/2024.20632

5.2. Deep Learning Model Construction

The core of the model is composed of a CNN and a MLP, which are used to process Mel spectrograms and global features extracted from CSV files, respectively. The CNN component consists of three convolutional layers, each using a different number of filters to extract various levels of features from the audio spectrum. Under each convolutional layer, a max-pooling layer is applied to downsample the features, helping to decrease the computational complexity and dimensionality of the model. Additionally, Batch Normalization is introduced after each convolution operation to accelerate the training process and improve the model's generalization capability. At the end of the convolutional layers, a flattening operation converts the high-dimensional features into a one-dimensional vector, which is then connected to the MLP network through fully connected layers.

5.3. Model Fusion and Classification

During the model fusion stage, the spectral features from the CNN and the global features processed by the MLP are concatenated into a comprehensive feature vector. In the MLP section, the input features are processed through two fully connected layers, using ReLU activation functions to enhance the model's nonlinear representation capacity. To prevent overfitting, Dropout is applied to both the CNN and MLP branches, randomly dropping some neurons to increase the model's robustness. The concatenated features are fed into the final fully connected layer, which uses a softmax activation function to get the predicted probabilities for 10 music genre categories, thus completing the classification task.

5.4. Model Training and Evaluation

The model is optimized using the Adam, with cross-entropy loss as the loss function. To get better training efficiency and model performance, the training is conducted for 150 epochs, with a batch size of 100 for batch training. During the training process, the model's performance on unseen data is continuously evaluated using the validation set to ensure effective generalization. After training is complete, the model is evaluated on the test part, and the this accuracy and loss values are reported, demonstrating the model's performance.

5.5. Results Saving and Visualization

After training and evaluation, the model saves the final training results, including the trained model parameters, features from the training and test sets, and label data. Additionally, the code generates a report and a confusion matrix to analyze the model's accuracy. By visualizing the classification results, researchers can clearly see which categories the model performs well on and where confusion occurs. To further find out the performance, the code also visualizes accuracy and loss curves from the training process.

6. Evaluation

The traditional music genre classification algorithms proposed by Tzanetakis and Cook, which utilized features like MFCCs combined with classifiers such as KNN and GMM, achieved an accuracy of only 61% on the GTZAN dataset [5]. In contrast, our dual-branch CNN and MLP fusion algorithm reached an accuracy of 77%, representing a significant improvement over traditional method.

Table 1. The image of the comparison

Researcher	George Tzanetakis,Perry Cook	Xiang Li, Fan Li.Ziyue Yang, Zhipeng Lu
Subject	Music genres classification	
Data Set	GTZAN	
Methods	KNN. GMM. SVM	CNN, MLP
Main features	Loudness Zero-Crossing,Rate MFCC Chroma	MFCC, Chroma, Mel Spectrum
Accuracy	61%	76%

DOI: 10.54254/2755-2721/132/2024.20632

7. Results

The results as follow:

- Precision: 0.77
- Recall: 0.67
- F1-Score: 0.65
- Support: 10

The images of the results as follow:

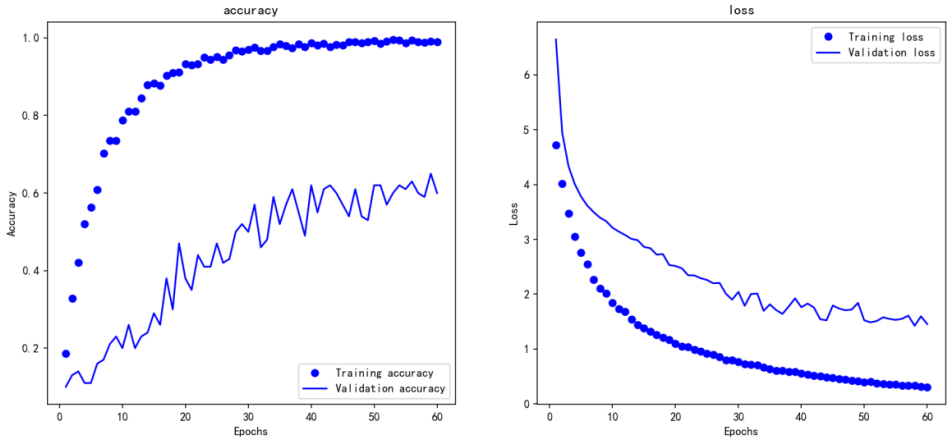

Figure 2. The chart of model training trend

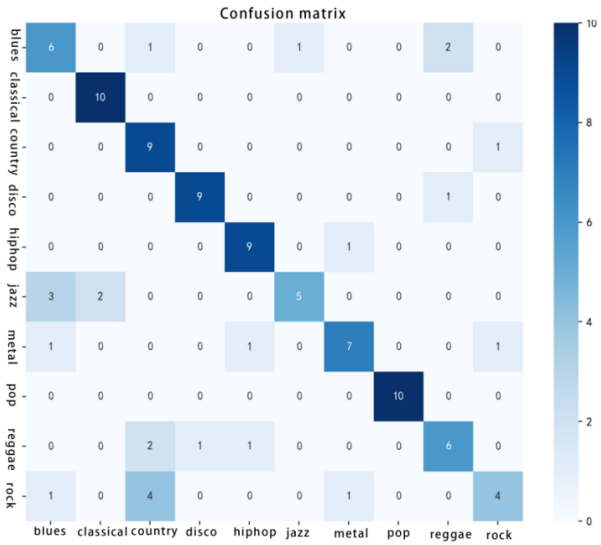

Figure 3. The image of Confusion matrix

8. Conclusion

8.1. Advantage

We combined a convolutional neural network (CNN) and a multi-layer perceptron (MLP), inputting the spectral features and global features of the corresponding audio respectively, and efficiently achieved music genre classification. In the preprocessing stage, we extracted features such as Mel spectrogram, chrominance, and MFCC for the model, which provided rich training parameters for the CNN model. A

DOI: 10.54254/2755-2721/132/2024.20632

rich set of global feature data is input into the MLP model for training. In the modeling stage, we use deep learning to help capture complex patterns in audio signals. Finally, by applying model fusion and regularization techniques, we successfully achieved the classification of music genres. At the same time, through visualization technology, we can also clearly see the performance of the classification model. Our method highlights the potential of deep learning in audio classification tasks and can also provide relevant references for future research in similar fields.

8.2. Limitation

8.2.1. Limited Data. In the training dataset, there are only 100 audio samples for each music style, and the data is slightly lacking. Therefore, the training of our model is limited by the diversity and scale of the dataset, and there is a problem of insufficient generalization ability for some niche music styles.Moreover, a small-scale dataset can also lead to overfitting, causing the perform poorly on unseen data.

Simplified Fusion Strategy: The feature concatenation strategy used by our model to fuse the outputs of the CNN and MLP is overly simplistic, which may not fully leverage the complementary nature of different types of features.

Risk of Overfitting: Due to the limited amount of data, even though we introduced Dropout and L2 regularization to prevent overfitting, the risk still exists. This may cause the model to perform extremely well on the training set but unsatisfactory on the validation, impacting the model's practical application.

Time Efficiency Issues: Our model uses a large number of convolutional layers and fully connected layers, which may require a long computation time and high computational resources during data preprocessing and training.

8.3. Implementation

8.3.1. Data Augmentation and Enhancement. In the future, more data should be added, and data should be enhanced. Data augmentation techniques can be used to expand the existing dataset, such as generating new training samples by changing the speed, pitch, time shift, or adding background noise to the audio. This can help achieve the goal of expanding the data.

Introduction of More Complex Fusion Mechanisms: Attempts should be made to use adaptive weighting and attention mechanisms, which can help the model achieve more intelligent combinations between different features. By better highlighting key features during fusion, the model is expected to improve classification accuracy.

Optimization of Model Structure and Regularization: Efforts should be made to simplify the model structure, reducing unnecessary convolutional and connected layers to lower the overfitting and speed up the training. More regularization techniques can be used, such as a higher Dropout ratio or more sophisticated methods like Early Stopping to prevent overfitting.

Application of More Efficient Training Methods: Consider adopting faster optimization algorithms or distributed computing frameworks, such as using TensorFlow's distributed training. Mixed precision training could also be introduced in the training process to accelerate model training while reducing the demand of high computational resources.

9. Critical Analysis and Future Work

While our proposed model shows a degree of innovation and achieved a 77% accuracy after iteration, there is still a noticeable gap compared to the 90% accuracy achieved by foremost methods. This indicates that our model still has considerable room for improvement.

In terms of the model, we will explore the fusion of additional model types and increase the depth and complexity of the model.

Regarding feature processing, we will explore more features and tailor them for model input, enhancing the diversity of features.

DOI: 10.54254/2755-2721/132/2024.20632

In terms of the dataset, we will seek to include music from various regions and cultures to train the model, increasing the richness and improving its generalizability.

Acknowledgement
Xiang Li, Fan Li, Zhi Peng Lu, and Zi Yue Yang contributed equally to this work and should be considered co-first authors.

References

[1] Wei Y, Xia W, Lin M, Huang J, Ni B, Dong J, Zhao Y, Yan S (2016) Hcp: A exible cnn framework for multi-label image classification. IEEE Trans Pattern Anal Mach Intell 38(9): 1901–1907
[2] Ciresan D, Meier U, Masci J, Gambardella ML, Schmidhuber J (2011) Flexible, high performance convolutional neural networks for image classification, in IJCAI Proceedings-International Joint Conference on Artificial Intelligence, 22:1237, Barcelona, Spain
[3] Arabi A, Lu G (2009) Enhanced polyphonic music genre classification using high level features, In Signal and Image Processing Applications (ICSIPA), 2009 IEEE International Conference on, pp 101–106, IEEE
[4] Yang S, Hai Long W, Lin L, Dong Mei P.(2022). A review of the classification of music genres in the field of music information retrieval. FX361, 30 June. Available at: https://m.fx361.com/news/2022/0630/18906073.html (Accessed: 16 August 2024)
[5] Tzanetakis, G. & Cook, P., 2002. Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, 10(5), pp.293-302. Available at: https://ieeexplore.ieee.org/document/1021071 [Accessed 18 Aug. 2024].
[6] Dieleman, S. & Schrauwen, B., 2014. End-to-end learning for music audio. arXiv preprint. Available at: https://arxiv.org/abs/1312.6614 [Accessed 18 Aug. 2024].
[7] Choi, K., Fazekas, G., Sandler, M. & Cho, K., 2017. Convolutional recurrent neural networks for music classification. arXiv preprint. Available at: https://arxiv.org/abs/1609.04243 [Accessed 18 Aug. 2024].
[8] Salamon, J. & Bello, J.P., 2017. Deep convolutional neural networks and data augmentation for environmental sound classification. arXiv preprint. Available at: https://arxiv.org/abs/1608.04363 [Accessed 18 Aug. 2024].
[9] Miron, M., Lattner, S. & Richard, G., 2019. A recurrent neural network approach to music genre classification using audio features. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Available at: https://ieeexplore.ieee.org/document/8744015 [Accessed 18 Aug. 2024].
[10] Fontes, A., Richard, G. & Peeters, G., 2022. Music genre classification using deep learning techniques. arXiv preprint. Available at: https://arxiv.org/abs/2107.03465 [Accessed 18 Aug. 2024].
[11] Anirudh Ghildiyal, Komal Singh, and Sachin Sharma, "Music genre Classification using Machine Learning", Proceedings of the 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA 2020) : 5-7, November 2020.
[12] Allamy S, Koerich A L. 1D CNN architectures for music genre classification[C]//2021 IEEE symposium series on computational intelligence (SSCI). IEEE, 2021: 01-07.
[13] D. Sun, M. Lv, H. Ren, J. Fan, and Q. Liu, "A Classification Model of Music Genres Using BP Neural Network Based on Genre Similarity Analysis, " in Proceeding - 2021 China Automation Congress, CAC 2021, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 6989–6994. doi:10.1109/CAC53003.2021.9728352.
[14] Elbir A, Aydin N. Music genre classification and music recommendation by using deep learning[J]. Electronics Letters, 2020, 56(12): 627-629.
[15] Xie C, Song H, Zhu H, et al. Music genre classification based on res-gated CNN and attention mechanism[J]. Multimedia Tools and Applications, 2024, 83(5): 13527-13542.