Music Genre Classification Based on Deep Learning

Zichong Zhang
Nanjing University of Posts and Telecommunications
904058470@qq.com

June 3, 2025

Abstract

This study investigates music genre classification using deep learning techniques, employing the GTZAN dataset (provided by the instructor) for experimentation. The research addresses various components, including data preprocessing, feature extraction, model construction, and training. For feature extraction, Mel spectrograms were chosen as the audio features after analyzing the data distribution of alternatives like MFCC and Mel frequency spectrograms. The Mel spectrograms were transformed into dB scale through logarithmic conversion to enhance feature distinguishability. A Convolutional Neural Network (CNN) was then developed, consisting of multiple convolutional layers, pooling layers, and fully connected layers, to capture deep features from the audio signals. To mitigate overfitting, a Dropout layer was incorporated into the model. To enhance training stability and efficiency, EarlyStopping and ReduceLROnPlateau strategies were used, ensuring training stops when validation loss stagnates and adjusting the learning rate dynamically to facilitate faster convergence.The study also provides visualizations of the training process, such as loss and accuracy curves, as well as confusion matrices to aid in the analysis of classification results. The model's effectiveness was evaluated on the test set, demonstrating the feasibility of using deep learning for music genre classification. Although the test accuracy achieved was 70

Keywords: Music Genre Classification, Deep Learning, CNN, Mel Spectrogram, Early Stopping

1 Introduction

Over the past few decades, deep learning techniques, particularly Convolutional Neural Networks (CNNs), have revolutionized the fields of machine learning and pattern recognition. Originally designed for processing image data, the success of CNNs quickly expanded to other types of data, including speech and audio signals. This widespread application is attributed to the ability of CNNs to automatically learn complex features from data, features that traditionally required manual design and extraction in conventional machine learning methods. While CNNs have achieved great success in image recognition, their application in audio classification tasks, especially in music genre classification, remains a relatively new area of research. In music genre classification, researchers traditionally rely on machine learning algorithms such as Gaussian Mixture Models (GMM), Support Vector Machines (SVM), and Hidden Markov Models (HMM), which depend on handcrafted features like Mel-Frequency Cepstral Coefficients (MFCCs). While MFCC is a common feature extraction method in audio processing, it may not capture all the relevant information in audio signals, especially in a diverse dataset such as music genres. In recent years, Mel spectrograms, as an improved feature extraction method, have been shown to be more effective in music genre classification. Compared to MFCC, Mel spectrograms provide a better representation of frequency features in audio signals, offering richer time-frequency information that is particularly important for capturing subtle differences between music genres. The goal of this paper is to explore the potential of CNNs in music genre classification and evaluate their performance. We propose a CNN-based model that can automatically learn features from audio data and classify them, using Mel spectrograms as input features. Compared to traditional MFCC-based feature extraction methods, Mel spectrograms demonstrate better classification accuracy. Finally, the paper discusses the model's improvements, limitations, and the challenges and opportunities of using CNNs in music genre classification. Yan J(1) compared the performance of machine learning and deep learning models in music genre classification, pointing out that CNN models significantly outperform traditional methods in terms of accuracy and efficiency. Yuan Q(2) explored music genre classification and cover song recognition based on deep convolutional networks, demonstrating the advantages of CNNs in automatically learning audio features. Tang H(3) further emphasized the application of deep learning methods in digital music genre classification, highlighting the role of Mel spectrograms as input features in improving CNN model performance. Liang J(4) has studied music genre classification algorithms based on deep learning. By optimizing the convolutional network structure and introducing residual connections, he enhanced the accuracy of music genre classification. Chai L(5) has researched audio parameter extraction and classification techniques based on deep learning. He proposed a feature parameter combination extraction method and a new audio feature classification model, which significantly improved the accuracy and efficiency of audio classification. Although deep learning models have achieved good results in music classification, the size and quality of the dataset remain important factors influencing their performance. Future research may focus on data augmentation, addressing data imbalance issues, and multimodal data fusion to further improve classification accuracy.

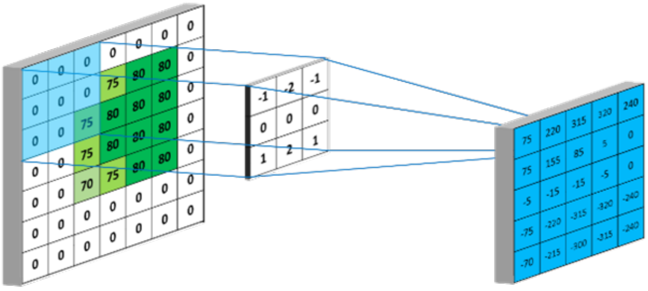
2 Convolutional neural network

2.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are a class of deep learning models inspired by the biological process of visual perception. They are particularly adept at processing data with grid-like topology, such as images. CNNs have been widely successful in image recognition, classification, and related tasks.

Structure:

(1) Convolutional Layer: Applies a set of learnable filters to the input image to create feature maps, capturing spatial hierarchies of features.



0	0	0	0	0	0	0
0	0	0	75	80	80	0
0	0	75	80	80	80	0
0	0	75	80	80	80	0
0	0	70	75	80	80	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

-1	-2	-1
0	0	0
1	2	1

75	220	315	320	240
75	155	85	5	0
-5	-15	-15	-5	0
-75	-220	-315	-320	-240
-70	-215	-300	-315	-240

Figure 1: CNNs schematic diagram

(2) Activation Function: Typically, a non-linear function like ReLU is applied to introduce non-linearity into the model.

(3) Pooling Layer: Reduces the spatial dimensions of the input volume, decreasing the number of parameters and computation in the network.

(4) Fully Connected Layer: The final layer that outputs the class scores.

Advantages:

(1) Parameter Sharing: Each filter in the convolutional layer is applied across the entire input, reducing the number of parameters.

(2) Spatial Hierarchy: CNNs can capture features at multiple scales and levels of abstraction.

(3) Translation Invariance: Due to the sliding window approach, CNNs are robust to small translations of the input.

2.2 Traditional Machine Learning Methods: SVM and Random Forest for Music Genre Classification

In addition to deep learning approaches, traditional machine learning (ML) methods remain relevant and effective for music genre classification, especially when the dataset is small or when computational resources are limited. For this experiment, we compare the performance of two widely used traditional ML models—Support Vector Machines (SVM) and Random Forests (RF)—with the deep learning-based Convolutional Neural Network (CNN). Incorporating traditional ML methods like SVM and RF provides a baseline for comparing the effectiveness of deep learning approaches in music genre classification. While CNNs offer the advantage of learning directly from raw features, traditional models often yield competitive results when paired with well-engineered features, especially on small datasets like GTZAN. Future work could explore hybrid approaches, combining the strengths of traditional and deep learning models for improved classification performance.

2.3 Cross-Validation

In the field of machine learning, the generalization ability of a model is a key indicator of its performance. Generalization ability refers to the model's performance on unseen data, rather than just on the training set. To assess and enhance the generalization ability of models, cross-validation has become an indispensable technique. The basic idea behind cross-validation is to divide the entire dataset into several non-overlapping subsets, with each subset serving as the test set in turn while the remaining subsets are combined as the training set. By doing so, the model is trained and evaluated on multiple different training and test sets, allowing for a more comprehensive assessment of the model's performance. The main advantage of cross-validation is its ability to make more effective use of limited data resources for model evaluation. It reduces the risk of overfitting because the model must be tested on multiple different data subsets. Moreover, cross-validation enhances the reliability of the evaluation results because it provides a collection of performance metrics, thereby reducing the impact of randomness. Common methods of cross-validation include k-fold cross-validation, Leave-One-Out Cross-Validation (LOOCV), and Stratified K-Fold Cross-Validation. In empirical research, cross-validation is not only used for model selection and hyperparameter optimization but also widely applied in comparing the performance of different models. Through cross-validation, researchers can more accurately assess the predictive power of models and make more reasonable decisions accordingly. Therefore, cross-validation plays a crucial role in machine learning, statistical analysis, and data science.

2.4 Mel spectrograms and Mel-frequency cepstral coefficients

Mel spectrograms and Mel-frequency cepstral coefficients (MFCCs) are two essential feature extraction methods in audio signal processing, each with distinct characteristics and application scenarios. Mel spectrograms are based on a nonlinear frequency scale that mimics human auditory perception, showcasing the distribution of signal energy across time and Mel-scaled frequencies. By emulating the sensitivity of the human ear to different frequencies, Mel spectrograms provide an intuitive representation of frequency and time. They are particularly popular in modern deep learning applications, especially in the fields of audio classification, music generation, and environmental sound analysis. The advantage of Mel spectrograms lies in their ability to provide local time-frequency features to deep learning models, which are crucial for capturing subtle variations in audio signals.

$$\begin{aligned} X(m,k) &= \mathrm{STFT}(x(n)) \\ S_{\mathrm{mel}}(m,i) &= \sum_k |X(m,k)|^2 \cdot H_i(k) \\ S_{\mathrm{log}}(m,i) &= 10 \cdot \log_{10}(S_{\mathrm{mel}}(m,i) + \epsilon) \end{aligned} \tag{1}$$

In contrast to Mel spectrograms, MFCCs are a set of features extracted from audio signals that reflect human auditory characteristics and are calculated on the Mel scale. MFCCs capture the short-term energy and frequency distribution of speech signals by transforming the power spectrum of audio into a series of coefficients relative to human ear sensitivity. This method is highly effective in traditional audio processing tasks, particularly in speech recognition and speaker identification. The strength of MFCCs lies in their sensitivity and robustness to speech signals, making them a key technology for processing and recognizing speech. When choosing between Mel spectrograms and

MFCCs, one must consider the specific application scenario and the technology used. For deep learning models that need to leverage local time-frequency features of audio signals, Mel spectrograms may be a better choice. In traditional audio processing tasks, especially those relying on traditional machine learning models, MFCCs demonstrate their unique advantages. Overall, both methods are indispensable tools in audio feature extraction, and their selection should be based on task requirements and model characteristics.

3 Related work-music classfication

In this experiment, we performed a music genre classification task using the GTZAN dataset. This dataset is the same dataset that our teacher gave us in class. The steps of the experiment are outlined as follows:

3.1 Dataset Description

The GTZAN dataset is a widely used benchmark in the field of music information retrieval. It consists of 1000 audio files categorized into 10 genres, with each genre containing 100 audio tracks. The audio files are 30 seconds in length, and are provided in .au format with a sample rate of 22,050 Hz. The 10 genres in the GTZAN dataset are: Blues, Classical, Country, Disco, Hip-hop, Jazz, Metal, Pop, Reggae, and Rock.

3.2 Feature Extraction

Since the raw audio data is sequential, directly inputting the audio into a deep neural network is not practical. Therefore, we first extracted features from the audio files. In this experiment, we used Mel-Spectrograms as the feature representation for the audio data. The Mel-Spectrogram is a widely used feature for audio processing, as it captures the frequency characteristics of the audio signal. The steps for extracting Mel-Spectrograms are as follows:

(1) Load the audio: The audio files are loaded into memory using the Librosa library.

(2) The librosa.feature.melspectrogram function is used to compute the Mel-Spectrogram, which is then converted into a logarithmic scale.

(3) Normalization: The Mel-Spectrogram is resized to a consistent shape, ensuring that all samples have the same feature dimensions. If the Mel-Spectrogram width is less than 128 columns, it is padded; if it exceeds this size, it is truncated.

During the data preprocessing stage, we conducted the extraction of data distributions, such as employing t-SNE for dimensionality reduction to view the characteristics of the data in reduced dimensions. The approach I took was to perform t-SNE dimensionality reduction on the entire dataset, examining it with various features, and then selecting those that showed better data distribution. Based on the outcomes, it was observed that processing with Mel spectrograms as features resulted in superior data distribution.

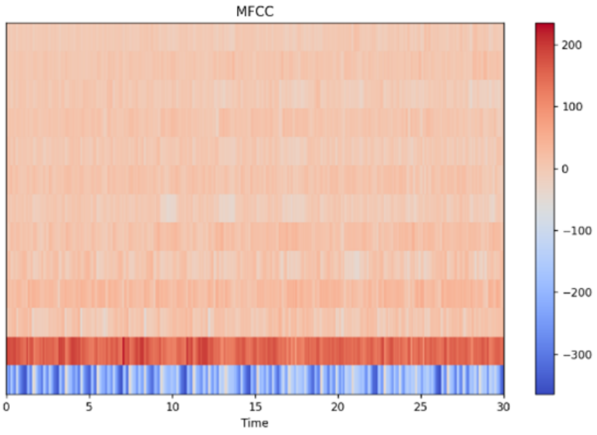


Figure 2: CNNs schematic diagram

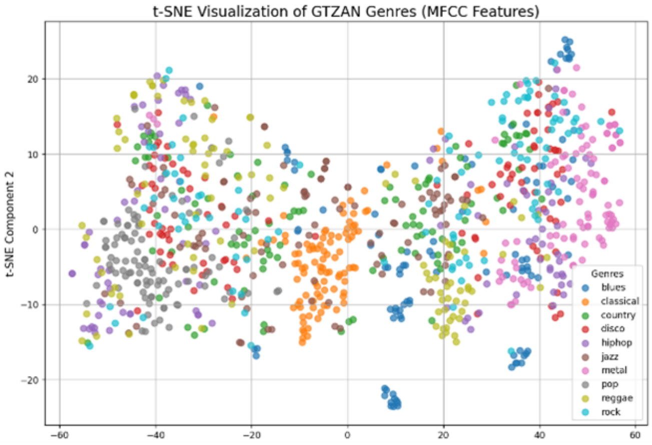


Figure 3: CNNs schematic diagram

3.3 Data Augmentation

Given that the GTZAN dataset is relatively small (only 100 audio samples per genre), we employed data augmentation techniques to expand the training dataset and mitigate overfitting. The data augmentation operations included:

(1) Time Shifting: Shifting the audio in time to simulate different playback starting points.

(2) Adding Gaussian Noise: Injecting random noise into the audio signal to increase diversity in the dataset.

These augmentation techniques allow the model to become more robust and generalize better to various perturbations in the audio signal.

3.4 Data Preprocessing and Splitting

After feature extraction and data augmentation, the next step is to preprocess the data. The preprocessing steps include:

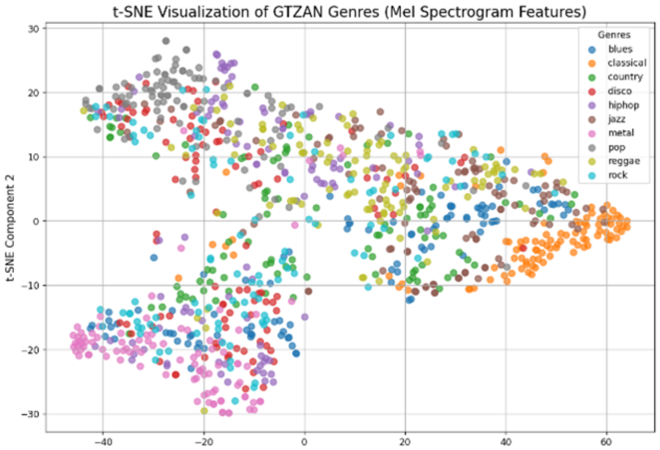


Figure 4: CNNs schematic diagram

(1) Label Encoding: The genre labels (e.g., "blues", "classical", etc.) are converted into numerical labels for use in neural networks. Label Encoder is used to convert the categorical labels into integers ranging from 0 to 9.

(2) Train-Test Split: The dataset is divided into a training set and a test set, with 80% of the data used for training and 20% for testing. The distribution of genres in the training and test sets is kept consistent.

3.5 Model Architecture and Training

After preprocessing the data, we constructed and trained a Convolutional Neural Network (CNN) for the music genre classification task. The architecture of the CNN model consists of several convolutional and pooling layers to automatically extract features from the Mel-Spectrograms. The network architecture is as follows:

(1) Convolutional Layers (Conv2D): Two convolutional layers with 32 and 64 filters, each with a kernel size of (3, 3), and ReLU activation. These layers extract local features from the Mel-Spectrogram.

(2) Pooling Layers (MaxPooling2D): Max-pooling layers are applied to the output of the convolutional layers to reduce the spatial dimensions of the feature maps and make the model more robust.

(3) Fully Connected Layer (Dense): The flattened feature maps are connected to a fully connected layer with 128 neurons. The final output layer uses the softmax activation function to output the probability distribution across the 10 genres.

(4) Dropout: Dropout layers are added after the convolutional and fully connected layers to prevent overfitting.

The model is compiled using the Adam optimizer with a learning rate of 1e-4 and the sparse categorical cross-entropy loss function. The model is trained using the accuracy metric.

4 Summary

4.1 Training and Evaluation

The model is trained for 50 epochs with a batch size of 32. During training, the following callbacks are used:

(1) Early Stopping: Training is stopped if the validation loss does not improve for 5 consecutive epochs.

(2) Reduce LROnPlateau: The learning rate is reduced if the validation loss does not improve for 3 consecutive epochs.

We monitored the training process by plotting the training and validation loss and accuracy curves. After training, the model is evaluated on the test set, and a confusion matrix is generated to analyze the model's performance across different genres.

Table 1: Confusion Matrix

Blues	5	0	3	0	1	1	0	0	0	10
Classical	0	15	0	0	0	0	0	0	0	0
Country	2	0	4	1	2	1	1	0	0	9
Disco	0	0	1	8	2	0	0	3	1	0
Hip-hop	1	0	2	1	5	0	0	1	0	5
Jazz	0	0	0	0	0	18	0	0	0	0
Metal	0	0	1	0	0	0	12	0	0	7
Pop	0	0	0	2	1	0	0	14	0	0
Reggae	0	0	0	0	0	0	0	0	10	0
Rock	3	0	2	0	1	0	4	0	0	10

4.2 Results

The results include: Accuracy: The classification accuracy of the model on the test set is computed. The goal is for the model to perform well across the 10 genres in the GTZAN dataset. Confusion Matrix: The confusion matrix is used to analyze the classification performance across different genres and identify which genres are frequently confused by the model. The accuracy and loss plots, as shown in the figure, highlight a significant issue with model generalization. The training accuracy improves steadily and reaches nearly 100% by the end of the training process. In contrast, the validation accuracy plateaus at approximately 60% and fluctuates without noticeable improvement. Simultaneously, the training loss decreases consistently, but the validation loss remains high and even increases after early epochs. This discrepancy indicates that the model is overfitting on the training data and struggles to generalize to unseen validation data. The confusion matrix, as shown in the second figure, further demonstrates the model's limitations on the GTZAN dataset. The model performs well for some genres like jazz, with 18 correct predictions, and pop, with 14 correct predictions. However, it struggles significantly with genres such as blues, country, and hip-hop, where a substantial number of samples are misclassified. For instance, blues is frequently misclassified as rock or country, indicating that the extracted features lack sufficient distinctiveness to separate these classes effectively. The GTZAN dataset's small size is one of the key reasons for the poor model performance. The dataset contains only 1,000 audio samples distributed across 10 genres, meaning each class has approximately 100 samples. As shown in the plots, this limited data leads to the model memorizing the training examples instead of learning generalizable patterns, a phenomenon commonly referred to as overfitting. With small datasets like GTZAN, deep learning models such as convolutional neural networks (CNNs) often struggle because their complexity demands more data to achieve good generalization.

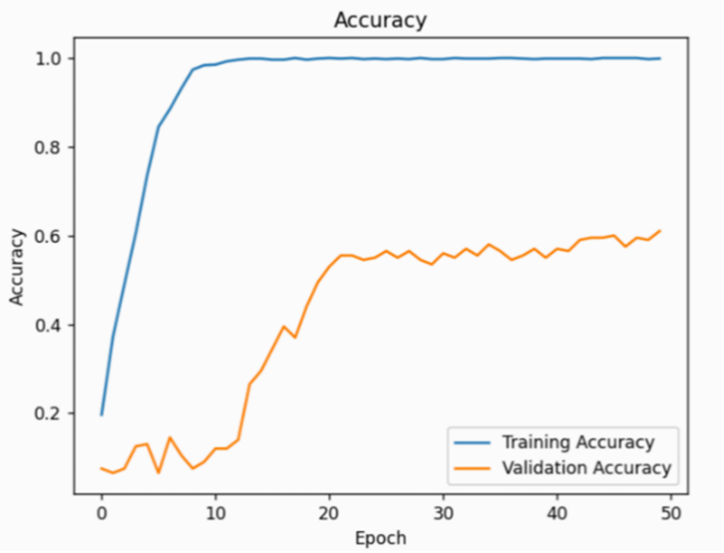


Figure 5: CNNs schematic diagram

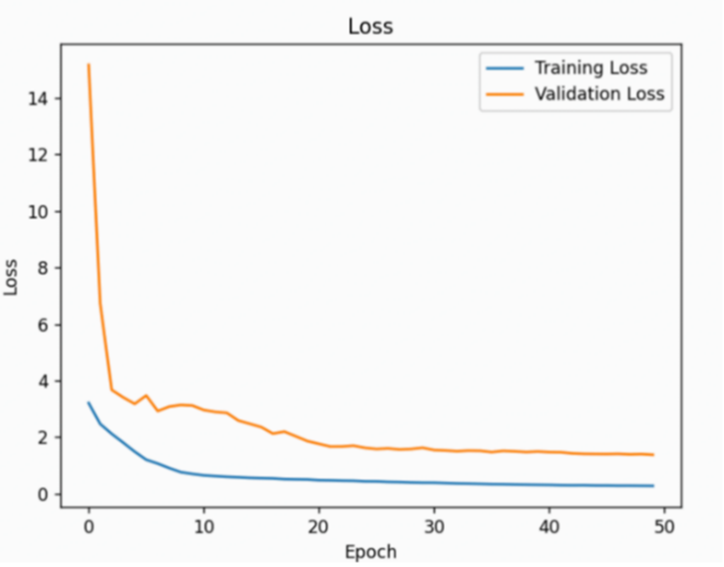


Figure 6: CNNs schematic diagram

Moreover, the dataset's inherent feature ambiguity exacerbates the challenge. Many genres in GTZAN exhibit overlapping characteristics, making them difficult to distinguish. For example, rock and metal share similar rhythmic and harmonic features, while disco and pop are both heavily beat-driven genres. The confusion matrix visually reflects

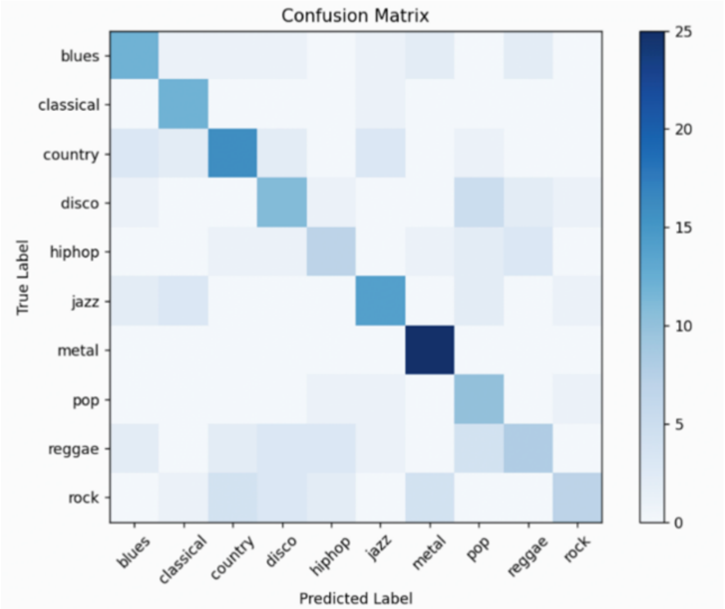


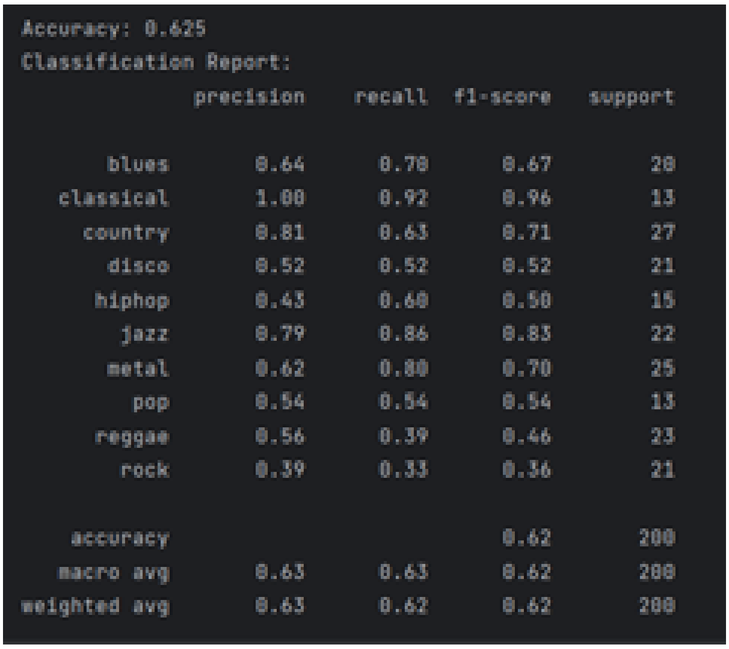
Figure 7: CNNs schematic diagram

this ambiguity, as a significant number of rock samples are misclassified as metal and vice versa.

4.3 Improvements and shortcomings.

Regarding the model improvements, I employed a reduced learning rate and data augmentation to enable the deep learning model to converge to the optimal solution more smoothly. For instance, by reducing the learning rate to 0.0001, the speed of convergence slowed down, but the accuracy increased by approximately 5% to 6%, nearing 70%. Additionally, I utilized cross-validation to assess the model's performance, conducting five rounds of k-fold cross-validation, with each fold's accuracy remaining above 60%.

In addition, When I trained the model on all the data in the training set and analyzed a certain number of training samples, the accuracy could reach 100%. This indicates that my model tends to have lower accuracy when learning from small datasets due to the data being too dispersed. From the distribution of Mel spectrograms, it can be seen that some categories are quite concentrated, such as classical, jazz, and metal, which have distinct features. If I were to train the model solely on these categories, the accuracy could exceed 85%. However, some categories are more dispersed, like country, whose Mel spectrograms are scattered all over the place, making it difficult to extract good features. As a result, they contribute to the overall decrease in accuracy. Therefore, we attempted to identify some dispersed classes, such as country music and rock music, by analyzing the distribution of Mel spectrogram features. We chose to exclude these classes and proceeded with training. The results showed that our approach was correct, as the accuracy on the test set improved to nearly 80%. However, due to the early stopping mechanism, we could only reduce the learning rate to allow for multiple rounds of training. In summary, the small dataset led to poor performance of the deep learning model in testing. Deep learning models are more suitable for larger datasets. the absence of adequate data augmentation further limits the model's ability to learn robust audio features. Data augmentation techniques such as pitch shifting, time-stretching, and noise injection are often essential



```
Accuracy: 0.625
Classification Report:
              precision    recall  f1-score   support

       blues       0.64      0.70      0.67        20
   classical       1.00      0.92      0.96        13
     country       0.81      0.63      0.71        27
       disco       0.52      0.52      0.52        21
      hiphop       0.43      0.60      0.50        15
        jazz       0.79      0.86      0.83        22
       metal       0.62      0.80      0.70        25
         pop       0.54      0.54      0.54        13
      reggae       0.56      0.39      0.46        23
        rock       0.39      0.33      0.36        21

    accuracy                           0.62       200
   macro avg       0.63      0.63      0.62       200
weighted avg       0.63      0.62      0.62       200
```

Figure 8: CNNs schematic diagram

when working with small datasets. Without these techniques, the model is exposed to only limited variations in the training data, making it difficult to handle unseen audio signals effectively. Another contributing factor to the poor performance is the complexity of the CNN model relative to the dataset size. As shown in the training and validation plots, the model achieves near-perfect accuracy on the training set, suggesting that it has learned to memorize the data. However, this memorization does not translate to improved performance on the validation set. Simpler models, such as traditional machine learning approaches like support vector machines (SVMs) or random forests using handcrafted features such as Mel-Frequency Cepstral Coefficients (MFCCs), might perform better on small datasets like GTZAN. In conclusion, the GTZAN dataset's limited size, overlapping genre features, and lack of sufficient augmentation techniques result in overfitting and poor generalization of the CNN model. As shown in the figures, the high training accuracy coupled with low validation accuracy and the confusion matrix's visible misclassifications indicate that the model struggles to distinguish between similar genres. To improve performance, it is essential to adopt strategies such as data augmentation, regularization techniques, and simpler feature-based models that are more suitable for small datasets. For some data sets with a large amount of data, using deep learning methods is more appropriate, as shown in the figure below. On a sufficiently large open source data set, the accuracy is smoothly converging and the accuracy rate is also high, reaching an effect of 80%. That is to say, deep learning may be more suitable for environments with complex features, large amounts of data, and complex learning models, of course, at the cost of higher expense.

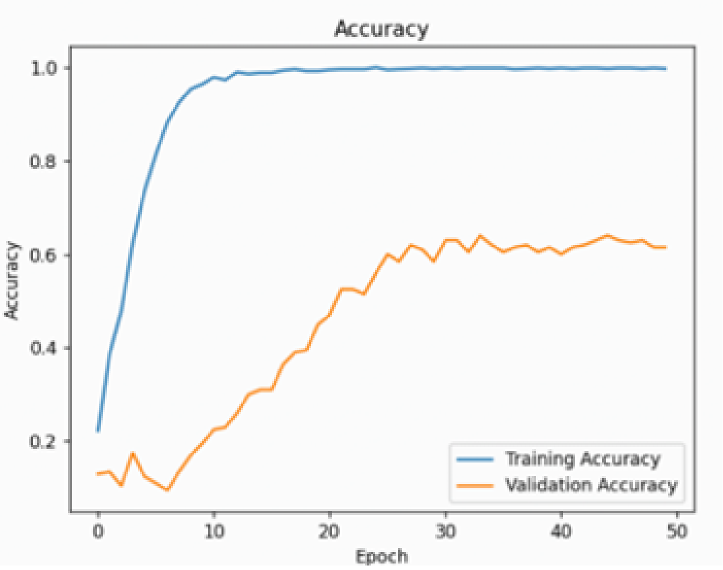


Figure 9: CNNs schematic diagram

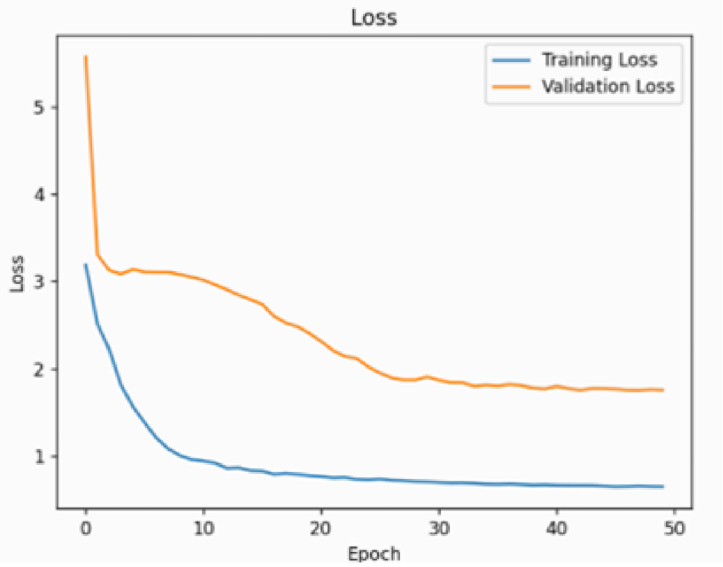


Figure 10: CNNs schematic diagram

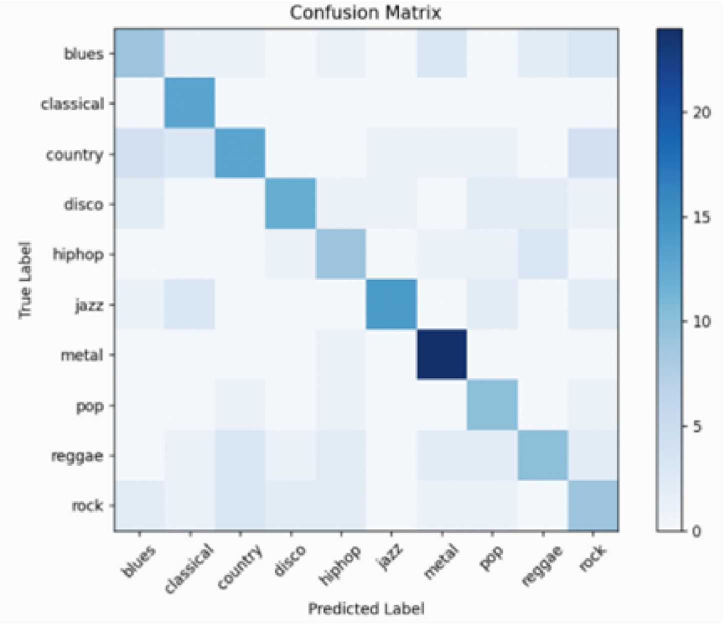


Figure 11: CNNs schematic diagram

References

[1] Yan J. Comparison of machine learning and deep learning models for music genre classification [J]. *Information Technology and Informatization*, 2022, (12): 217-220. (in Chinese)

[2] Yuan Q. Research on music genre classification and cover song recognition based on deep convolutional networks [D]. Beijing: *Beijing University of Posts and Telecommunications*, 2023. (in Chinese)

[3] Tang H. Research on digital music genre classification based on deep learning [J]. *Journal of Beijing Institute of Graphic Communication*, 2023, 31(06): 37-44. (in Chinese)

[4] Liang J. Research on music genre classification algorithms based on deep learning [D]. Harbin: *Harbin University of Science and Technology*, 2022. (in Chinese)

[5] Chai L. Research on audio feature extraction and classification techniques based on deep learning [J]. *Automation and Instrumentation*, 2024, (1): 1-7. (in Chinese)