Data augmentation techniques for classification of music genre

Aixin Zhang[1,3,*] **and Hanyu Zhang**[2,4,*]

[1]Westmount Collegiate Institute, 1000 New Westminster Dr, Thornhill, ON, L4J 8G3, Canada
[2]Zhengzhou No.7 Middle School Senior High School Branch, 70 Sanquan Rd, Jinshui District, Zhengzhou, Henan, China, 450045

[3]Joshuazhangax@outlook.com
[4]3067816183@qq.com
* These authors contributed equally to this work and should be considered co-first authors

Abstract. Music genre classification, a crucial aspect in the field of audio processing, has been widely investigated using different machine learning methods. However, there is still a need to determine the most optimal application between Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs) in the field of music genre classification, and the augmentation and enhancement of the data utilized for these machine learning techniques is still an important topic to study. The objective of this research is to identify the most effective solution and to improve the accuracy of music categorization. This will be achieved through an in-depth comparison of various data augmentation techniques and neural networks. In this study, both CNN and LSTM networks are optimized and trained. The dataset has been enhanced by extracting Mel Frequency Cepstral Coefficients (MFCC) data from music recordings using three different sampling techniques. The thorough approach undertaken facilitates a detailed assessment of the effectiveness of each method in classifying genres. The experimental findings show that the combination of CNN and random sampling outperforms all other tested algorithms, resulting in a significant enhancement in genre classification accuracy. This research provides valuable findings that can inform future studies in the pursuit of more effective techniques for classifying music genres.

Keywords: music genre classification, data augmentation, mel frequency cepstral coefficients

1. Introduction

The rapid growth of the internet has brought significant changes in various aspects of daily life, including how people interact with music. With the convenience of online music platforms, an increasing number of individuals now turn to music as a means of relaxation and enjoyment. Music has a long history of serving as a medium for expressing emotions, often through the use of musical instruments, either individually or within a group setting.

As internet develops and online music platforms gaining popularity, the traditional methods of offline music consumption have transitioned to digital platforms. This shift has resulted in substantial financial benefits for the music industry, with significant revenue growth reported by Chinese music platforms in 2018, witnessing a remarkable 20% increase in a single year, platforms have collectively amassed 180 million monthly users as of the year 2020 [1]. The shift is especially noteworthy in the post-Covid era, as premium music streaming platforms like Spotify and Apple Music have observed a rise in annual subscriptions [2].

To enhance user experience, music platforms continuously strive to improve their services, particularly in the areas of song selection and recommendation. Consequentially, research efforts in mass music information retrieval have increased significantly. A major challenge in this domain lies in efficiently organizing a vast repository of songs into meaningful categories. Traditionally, this task relied heavily on human resources, involving song tagging based on artist and album details and subjective categorization using listener preferences. However, this conventional approach had several inherent flaws, including potential biases from multiple listener impressions and susceptibility to malicious commentary, resulting in distorted outcomes. Moreover, this method demanded considerable allocation of financial, material, and human resources. Additionally, it often failed to consider crucial musical characteristics, such as rhythm, melody, and timbre, which could be utilized to create distinct music categories.

The classification of music according to its emotional content is a complex and important area of research. This is because the way people interpret music can differ subjectively, influenced by factors such as cultural upbringing and personal experiences. In order to unify music categorization, the utilization of computerized classification systems presents itself as a promising solution. These systems have the potential to address the limitations of traditional methods and enhance the accuracy and efficiency of music genre classification. Data augmentation techniques play a vital role in enhancing the accuracy and robustness of systems for music genre classification. By increasing the diversity and complexity of the available data, these techniques enable the classification model to effectively learn from a wider range of music samples. This article examines different methods of data augmentation for music genre classification. Its goal is to offer useful insights to researchers and professionals working for discerning music information and classification of music genres.

2. Literature review

There has been significant international interest in the classification of emotional content in music. Advancements in computer technology and the internet have enabled researchers to devise more effective methodologies, thereby improving the overall music access experience for listeners.

In a paper authored by Gao Yuxuan in 2020, the audio signals of music were represented using spectrograms, a methodology that effectively eliminates the potential errors associated with manual feature selection. Moreover, to enhance the music data for better training results, the researcher synergistically integrated signal characteristics with advanced data augmentation techniques, improving in both the accuracy and efficiency of the classification process [3].

Tan Zhiyuan's presented another solution in a 2020 research paper, an long short term memory (LSTM) based algorithm was employed for song emotion classification and investigated the use of the BERT model for lyric emotion classification. Additionally, he introduced an emotion lexicon to achieve balance and optimization in the lyric segment. By taking both lyrics and audio into account, his study exhibited a noteworthy improvement compared to prior works [1].

A feedback-based solution is covered in Zhong Huihu’s paper in 2021, by utilizing user’s comments collected from music platforms. After eliminating duplicate comments, over ten thousand data points were utilized. The researcher employed the LDA (Latent Dirichlet Allocation) machine learning model and the Snownlp library for text and sentiment analysis. This approach enabled a deeper understanding of the emotional perceptions of the audience towards songs, and the resulting sentiment keywords could be used as tags when launching the songs on the platform [4].

A frensh solution based on bottom-up broadcast neural network is presented in an article composed by Caifeng L, Lin F, aifeng L, Lin F, Guochao L, Huibing W, and Shenglan . The proposed method is called Bottom-up Broadcast Neural Network (BBNN), which is designed to effectively use both detailed and general information from various music to make better decisions. The highlight of their network architecture is that it’s designed to preserve the detailed information throughout the network, which is crucial for the task of music classification. The results on multiple datasets shows high accuracy and great promise [5].

A comparative study between classic ML methods and deep learning has been done by Dhevan L’s paper in 2020. The results indicate that both deep learning models and traditional classifiers, namely convolutional neural networks (CNNs), possess comparable capabilities in music genre classification. And slicing audio samples into smaller pieces improves classification precision [6].

Interestingly, a research lead by R Brzeski, M Cisynski, and D Kostrzewa presents a hybrid deep learning model in 2022. The proposed model combines CNN and LSTM networks together ensuring effective capture of temporal as well as spectral features in audio signals. The CNN network serves for extracting spectral characteristics at high-level, while the LSTM network serves to capture the temporal dynamics from the music. The findings indicate that the hybrid model proposed in this study is successful in capturing both spectral and temporal information for music genre classification. The model's high accuracy suggests that hybrid deep learning techniques have great potential in the field of music information retrieval [7].

These research endeavors to classify music based on expand on Musikalkemist’s music classification kit, by optimizing neural network architectures, utilizing multiple data augmenting techniques, and provide an in-depth comparison and analyzation on several approaches.

3. Method

3.1. Optimization

In this project, both CNN and LSTM are utilized. Since MFCC data is two dimensional, it is feasible to employ the data into convolutional neural networks. To suppress overfitting, two different methods are used.

3.1.1. Dropout. Dropout is a regularization technique employed during training to randomly deactivate certain neurons, so a different combination of neurons is used to predict the final output every iteration. By doing so, the model becomes less reliant on specific neurons, leading to improved generalization [8]. In both CNN and LSTM, dropout rates of 0.3 are chosen, meaning each neuron has a 30% chance to be shut down each iteration.

3.1.2. Ten-fold cross-validation. Ten-fold cross-validation is a commonly used evaluation technique in the field of machine learning. The dataset is partitioned into ten equally sized subsets, commonly referred to as "folds." The model is then trained and assessed ten times, with each iteration utilizing a different combination of nine folds for training and a single fold for testing. This helps assess the model's generalization capabilities and detect potential overfitting [9]. Ten fold cross validation provides a balanced trade-off between efficiency and statistical robustness in evaluating a model's performance.

3.2. LSTM approach

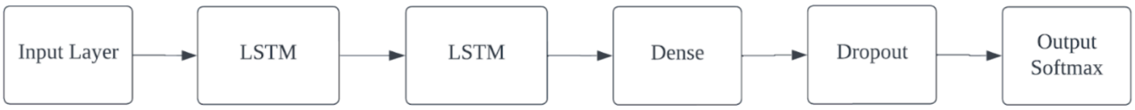


Figure 1. Long short memory architecture.

The Long Short-Term Memory is a specific type of Recurrent Neural Network (RNN) layer that is designed to address the challenge of learning long-term dependencies. Unlike traditional RNNs, LSTMs are able to overcome the "vanishing gradient problem" that often hinders the ability of RNNs to capture long-term dependencies. This LSTM model includes two layers, each consisting of 64 units. The first layer is set to return sequences, meaning it returns its full sequence of outputs for each sample at each timestep, rather than just the output at the final timestep. This is necessary when stacking LSTM layers, as the subsequent layer expects input in the same format.

Following the LSTM layers, a dense layer with 64 neurons is added, incorporating the Rectified Linear Unit (ReLU) as the activation function, injecting a degree of non-linearity into the model. The ReLU function has the advantage of helping the model learn complex patterns by allowing for a non-linear transformation of the input, avoiding problems of neutrons becoming stuck and eventually become unbailable, which may occur in other activation functions [10].

In order to strike a balance between model resilience and prediction accuracy, a dropout layer with a rate of 0.3 is implemented to mitigate the issue of overfitting. Concurrently, the output layer generates probabilistic multi-class predictions using the 'softmax' activation function. These final steps of the algorithm provide fast learning and accurate forecasting.

3.3. CNN approach

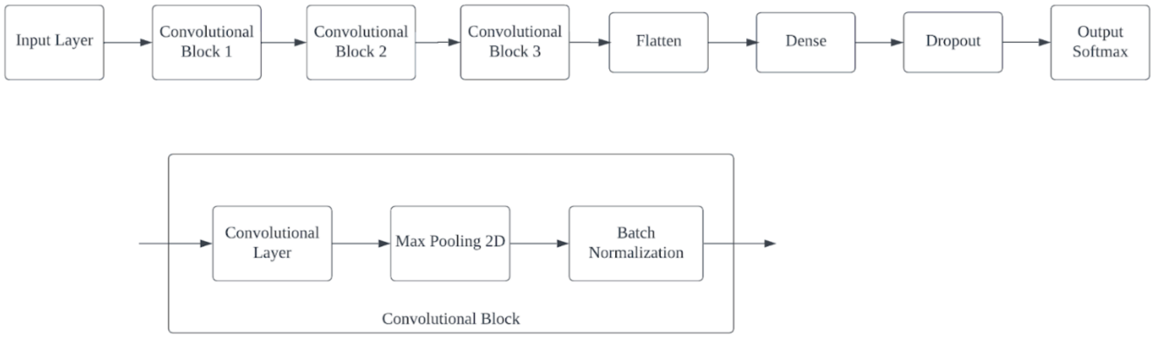


Figure 2. Convolutional neural network & convolutional block architecture.

Tensorflow Keras is used to construct the CNN architecture, three convolutional blocks are used each with a conv2d layer, a max pooling layer, and a batch normalization layer. Max pooling is a down-sampling technique, where the input 2D array is divided into non-overlapping regions, and the maximum value within each region is selected to create a smaller, more abstract representation of the original image, aiding in feature extraction and reducing spatial dimensions. While batch normalization normalizes the activations of each layer's neurons across a mini-batch of data, ensuring stable training by reducing internal covariate shift, improving convergence, and enabling higher learning rates.

In the convolutional layer, 32 filters are included, each with a size of 3x3. The activation function used for each neuron in this layer also are Rectified Linear Unit (ReLU).

In the max pooling layer, a pool size of 3x3 and a stride of (2, 2) is chosen, meaning the pooling window moves two pixels horizontally and two pixels vertically at each step. "padding='same'" argument is applied to ensure the output size remains the same as the input size after applying the pooling operation.

The CNN architecture uses the same dropout and activation layer as the LSTM version. However, a Flatten() layer is added beforehand to transform the tensor into one dimensional.

4. Experiment

4.1. Dataset

In this project, GTZAN Dataset on Kaggle is used. This dataset contains 30 second audio clips from the famous GTZAN dataset, and is separated into 10 different genres with 100 audio files each. However, some broken files were spotted in the database, thus a total of 6 audio files were removed from the original collection [11] [12].

Table 1. Sample count for each genre.

Genre	Blues	Classical	Country	Disco	Hip Hop	Jazz	Metal	Pop	reggae	rock
Number of samples	100	100	100	95	100	99	100	100	100	100

4.2. Data sampling

In order to find out the optimum sampling method for this dataset, three different types of sampling techniques are used, namely sequential sampling, random sampling and overlap sampling. In order to maintain the consistency of each audio clip, the sampling methods break the original 30 second audio into numerous 3 second segments. The random sampling and overlap sampling method splits every audio file into 20 clips, while overlap sampling samples 19 samples, each overlapping 50% with the previous one.

4.3. MFCC generation

Mel Frequency Cepstral Coefficients (MFCC) is a spectral representation technique used to emphasize audio characteristics that are significant for human auditory perception. This process is highly effective in extracting pertinent features from intricate sounds, and it has been empirically demonstrated as a reliable solution to distinguish various music genres [13].

4.3.1. Steps in MFCC.

- The sampling techniques are first applied, using the 3 second segments for further processing
- A pre-emphasis filter is applied to boost the high frequencies in the signal. This phase aims to equalize the frequency spectrum by adjusting the magnitudes of high frequencies, which tend to be lower compared to lower frequencies. Additionally, it helps prevent any numerical issues that may arise during the Fourier transform process.
- The signal is then divided into short frames because audio signals are non-stationary, meaning their frequency content changes over time. By framing, we can analyze those changes.
- Next, the Fourier Transform is computed to convert the signal from the time domain to the frequency domain. This enables further examination of the frequency components of the audio signal.
- The power spectrum calculated of each frame is then is calculated from the Fourier Transform. This gives us the power (or energy) at each frequency, which is a fundamental aspect of the sound.
- The Mel filterbank is subsequently utilized to replicate the nonlinear auditory perception of sound in humans. The Mel scale is a method used to establish the central frequencies of the filters in the Mel filterbank. It is defined as:

$$m = 2595 \cdot \log10(1 + 700f)$$

- The Mel filterbank consists of multiple filters, each specifically designed to detect and respond to a distinct range of frequencies. The filters' centre frequencies are determined using the Mel scale, which ensures that they are evenly spaced in terms of human auditory perception rather than Hertz.
- The logarithm of each Mel spectrum coefficient is then taken to mimic the human perception of loudness, which is logarithmic instead of linear [14].

- The Discrete Cosine Transform (DCT) is used to remove the correlation between the log Mel spectrum coefficients. As a result, a collection of MFCCs is obtained, which exhibits a reduced correlation compared to the log Mel spectrum coefficients. The process can be presented as:

$$Ci(n) = \sum_{m=0}^{M-1} Li(m)\cos(2M\pi n(2m+1))$$

4.3.2. Parameters for MFCC. In this study, audio is converted into MFCC signals for neural networks to use. The function librosa.feature.mfcc() is used for generation, with the parameters stated as below [15]:

Table 2. Value for each argument used in the librosa.feature.mfcc() function.

Parameter	Value
Sample rate (sr):	22050
Number of MFCCs (num_mfcc):	13
Window Size (n_fft):	2048
Samples between frames:	512

5. Result

In order to evaluate the models, two metrics are applied in stages:

Accuracy: The proportion of correctly identified test samples.

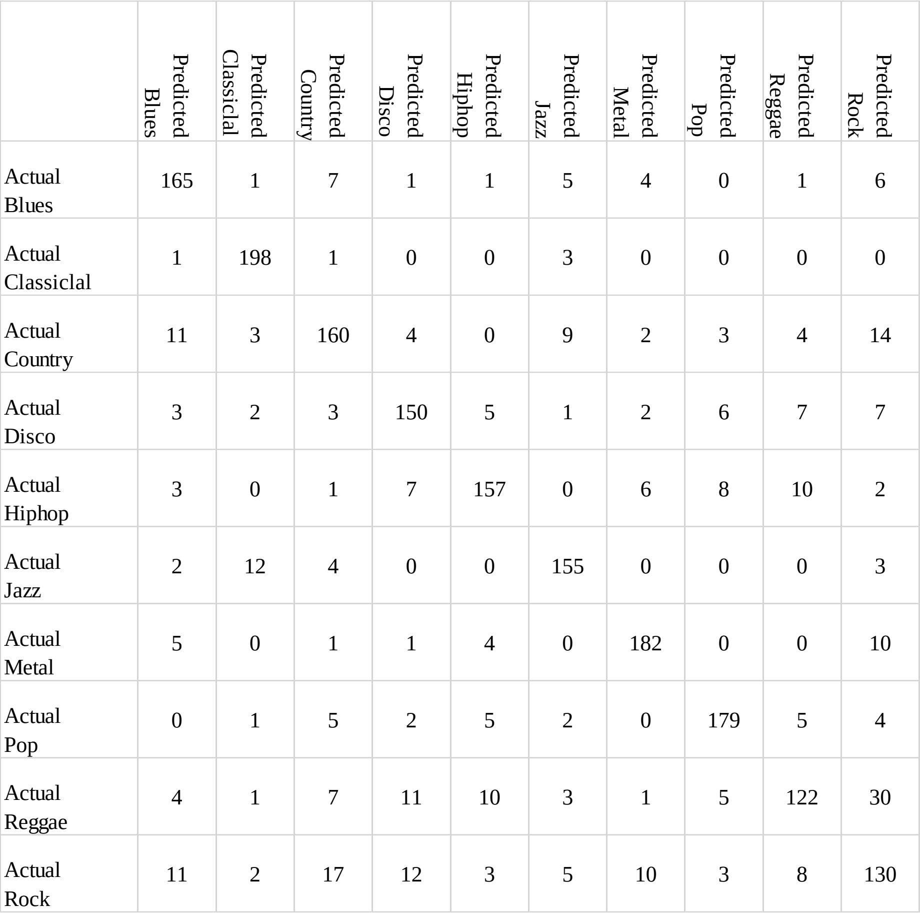
Confusion matrix: A statistical tool used in classification tasks to visualize a model's performance. It presents the count of correct and incorrect predictions, classified by each category, presenting the performance of the model in a per-category way.

Taking advantage of 10-fold cross validation, the mean of ten results are used as the final evaluation data.

Table 3. mean accuracy for every model/sampler in percentage.

	CNN	LSTM
Sequential Sampler	77.75	72.83
Overlap Sampler	81.30	75.20
Random Sampler	82.90	78.07

As for the confusion matrix, it is plotted for CNN architecture using the random sampling method, as it has the best performance thus most appropriate for discussing the details of the algorithm.



	Predicted Blues	Predicted Classiclal	Predicted Country	Predicted Disco	Predicted Hiphop	Predicted Jazz	Predicted Metal	Predicted Pop	Predicted Reggae	Predicted Rock
Actual Blues	165	1	7	1	1	5	4	0	1	6
Actual Classiclal	1	198	1	0	0	3	0	0	0	0
Actual Country	11	3	160	4	0	9	2	3	4	14
Actual Disco	3	2	3	150	5	1	2	6	7	7
Actual Hiphop	3	0	1	7	157	0	6	8	10	2
Actual Jazz	2	12	4	0	0	155	0	0	0	3
Actual Metal	5	0	1	1	4	0	182	0	0	10
Actual Pop	0	1	5	2	5	2	0	179	5	4
Actual Reggae	4	1	7	11	10	3	1	5	122	30
Actual Rock	11	2	17	12	3	5	10	3	8	130

Figure 3. Confusion Matrix of CNN architecture and random sampler.

Referring to the confusion matrix, this model exhibited high accuracy in classifying classical, metal and pop. The model isn’t too effective at judging reggae and rock, often misclassifying them to other genres. The rest of the tags has relatively promising results. The sample counts across different classes were also well balanced, which minimizes class bias and promoting the model's overall robustness and predictive accuracy.

6. Conclusion

This study focused on the classification of music genres using the GTZAN Dataset and investigated data augmentation techniques to enhance the classifiers' performance. The audio signals were analyzed by extracting Mel-frequency cepstral coefficients. Two neural network architectures, namely CNN and LSTM, were utilized for this purpose. Three different sampling methods were employed in order to train the classifiers. Cross-validation and dropout techniques were employed for optimization, revealing that the CNN-based model demonstrated superior performance compared to the LSTM model, making it more suitable for music genre classification. A comparison of different sampling approaches revealed that random sampling proved to be more effective than sequential and overlap sampling methods. Using CNN with Random sampling exhibited relatively promising outcomes, with a strong ability to accurately classify classical, metal, and pop genres. Nevertheless, it faced difficulties in effectively categorizing rock and country genres, frequently misidentifying them as alternative genres. For future research, exploring various pre-processing techniques could enhance data reliability. Additionally, modifying the sampling process could improve model robustness, and expanding the genre class counts might contribute to a more comprehensive understanding of related studies.

References

[1] Tan, Zhiyuan. (2020). Research on Multimodal Music Emotion Classification of Audio-Text Fusion. North China University of Technology.

[2] Denk, J., Burmester, A., Kandziora, M., & Clement, M. (2022). The impact of COVID-19 on music consumption and music spending. Plos one, 17(5), e0267640.

[3] Gao, Yuxuan. (2020). Research on Music Audio Classification based on Deep Learning. South China University of Technology.

[4] Zhong, Huihu. (2022). Research on Sentiment Classification of Pop Music App Reviews. Guilin University of Electronic Technology.

[5] Liu, C., Feng, L., Liu, G., Wang, H., & Liu, S. (2021). Bottom-up broadcast neural network for music genre classification. Multimedia Tools and Applications, 80, 7313-7331.

[6] Lau, Dhevan & Ajoodha, Ritesh. (2022). Music Genre Classification: A Comparative Study Between Deep Learning and Traditional Machine Learning Approaches. 10.1007/978-981-16-2102-4_22.

[7] Kostrzewa D, Ciszynski M, Brzeski R. Evolvable hybrid ensembles for musical genre classification[C]//Proceedings of the Genetic and Evolutionary Computation Conference Companion. 2022: 252-255.

[8] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

[9] Berrar, Daniel. (2018). Cross-Validation. 10.1016/B978-0-12-809633-8.20349-X.

[10] Agarap, A. F. (2018). Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375.

[11] Sturm, B. L. (2012, November). An analysis of the GTZAN music genre dataset. In Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies (pp. 7-12).

[12] GTZAN, https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification

[13] Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. IEEE Transactions on speech and audio processing, 10(5), 293-302.

[14] Patil, N. M., & Nemade, M. U. (2017). Music genre classification using MFCC, K-NN and SVM classifier. International Journal of Computer Engineering In Research Trends, 4(2), 43-47.

[15] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015, July). librosa: Audio and music signal analysis in python. In Proceedings of the 14th python in science conference (Vol. 8, pp. 18-25).