Musical Genre Classification

Joey Zheng
joeyz@mit.edu

Daniel Kim
dyk0518@mit.edu

A. Abstract

Musical Genre Classification (MGC) is the process that predicts a musical genre given a musical audio input. Previous work has used both visual and audio features representation of the input, achieving as high as 93.4% accuracy on the GZTAN dataset. When we limit the scope of feature extraction to only visual representation, the best accuracy achieved by the previously published models is 89.30%. In this paper, we show that higher accuracy can be achieved by extracting more and combining multiple audio-visual features, as well as making modifications to the model architecture. By using a combined audio-visual feature of Mel-spectrogram, MFCCs, and Chromagram with a hybrid CNN + Bi-GRU architecture, we achieved an empirical accuracy of 89.94%. This accuracy is higher than the accuracy achieved by other published models that conduct MCG on the GTZAN dataset using audio-visual features only.

B. Introduction

Think of your favorite genre of music. Have you ever wondered, what it is about that genre that you specifically enjoy? Whether it be improvisation in jazz or simple harmonies in country, each genre has specific musical features that attract its audience. With recent signal processing and machine learning development, we can computationally extract and identify these key audio features from musical pieces [11].

In this paper, we aim to use these extracted features from musical audio and use them to predict the musical genre. This process is known as Musical Genre Classification (MGC). With countless musical pieces produced every day in the prolific music industry, automating the classification and analysis tasks can benefit the field as a whole. There are many applications for this concept. Digital music services, such as Spotify, use MGC in the process of music categorization and recommendation, as well as providing data sources that could be analyzed through MGC in third-party research [3]. Furthermore, MGC helps expand and develop the broader concept of Music Information Retrieval (MIR), a multidisciplinary field that focuses on the extraction, analysis, organization, and retrieval of music-related information [4]. In this paper, we explore the current state of MGC using visual features, and how we improved it with data augmentation and architecture modifications of the model.

C. Background and Related Work

The standard dataset that is used for MGC is the GZTAN dataset, which has become the benchmark for musical analysis [13]. Hence, we also conduct our study using the GZTAN dataset. Currently, one of the best-performing MGC models for this dataset was introduced by Dai et al., which uses Mel-Frequence Cepstral Coefficients (MFCCs) and other audio features as the inputs with Deep Neural Networks to achieve 93.4% accuracy [2]. Dai et al. use two separate pipelines, Visual Feature Extraction (VFE) and Audio Feature Extraction (AFE) modules, during its data feature extraction process.

In our paper, we narrow down the scope to the MGC models that focus on using the image-level feature extraction of the musical audio input. By excluding the audio feature extraction in the process, we limit our study to how well computer vision techniques can be used for MGC. Within this narrowed scope, the best-performing model has been published by Ashraf et al. [1]. In their study, Ashraf et al. compare the performance based on two visual features from the audio: Mel-spectrogram and MFCCs, as well as a hybrid architecture of CNN and variants of RNN such as LSTM, Bi-LSTM, GRU, and Bi-GRU. Empirically, the best accuracy of 89.30% was achieved through the proposed hybrid architecture of CNN and Bi-GRU using Mel-spectrogram.

There are, however, a few ways that we suggest could improve this MGC using audio-visual features. We highlight two main ways this improvement could be achieved, data augmentation and architecture modification, which are also the main contributions of our paper:

- In addition to two audio-visual features used in the previous state-of-the-art model, Mel-spectrogram and MFCCs, we also use Chromagrams. Moreover, rather than training the model using each audio-visual feature individually, we also try training the model with combined audio-visual features. This way, each audio sample is represented with more diverse, informative features in the model.
- Compared to the previous state-of-the-art model, we reduce the number of CNN layer blocks, which helps with the computational efficiency of the model. We also replace the original RNN layer, which consisted of GRU, Bi-GRU, and GRU, with two Bi-GRU layers that led to better performance of the model.

With these changes, our model achieved the best accuracy of 89.94%, surpassing the performance of the previous

MGC models that use audio-visual features. In the following sections, we dive deeper into our methodology.

D. Methodology

D.1. Hypothesis

Our hypothesis is that a musical genre classification model that trains using combined audio-visual features will outperform those that train using only one audio-visual feature. We base our hypothesis on two key observations. Firstly, once audio-visual features are extracted, they can be simply treated as images. This implies that image classification techniques, such as CNNs, can be applied to solve this classification problem. Secondly, we observe that audio is multi-dimensional. One could extract many features from a piece of audio (especially musical audio) such as frequency, timbre, rhythm, dynamic, etc. Previous work in musical genre classification tends to focus on one audio-visual feature at a time. We want to investigate the potential of combining audio-visual features for training musical genre classification models.

D.2. Dataset

Our model was trained on the GZTAN dataset. This is a publicly available dataset consisting of 10 labeled genres: blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, and rock. Each genre comes with 100 audio tracks each 30 seconds in length. The audio tracks are all 22050Hz mono 16-bit audio files in .wav format with a storage size of about 1.3MB. These audio files were collected in 2000-2001 from various recording conditions such as personal CDs, radio, microphone recordings, etc. Additionally, the dataset includes 2 CSV files with statistics on the full audio file (30 seconds) and split audio file (3 seconds). Among some of the statistics are spectral centroid means, spectral centroid variances, MFCC means, MFCC variances, tempo, and more. However, we did not utilize these statistics, which we will discuss in the future work section. We chose the GZTAN dataset as it serves as the musical analysis benchmark in the field of musical genre classification.

D.3. Audio-Visual Features

Audio-visual features refer to features extracted from the audio that has visual components and can be analyzed with computer vision techniques. More specifically, this paper works with features such as spectrograms, which span the time axis and frequency axis with the frequency intensity at each point. This is in contrast with features such as spectral centroids, which measure the weighted mean of the frequencies at a given time.

This paper utilizes three audio-visual features: Mel-spectrograms, MFCCs, and Chromagrams.

D.3.1 Mel-spectrogram

Spectrograms have high utility for visualizing audio. Audio can be thought of as a combination of varying amplitudes of frequency over time. A spectrogram decomposes audio into its time and frequency components. Specifically, it maps a given time and frequency to the intensity of that frequency at that time. This is achieved by applying the Short-Time Fourier Transform (STFT). The STFT is a Fourier transform performed on smaller windows, or segments, of the audio. This allows for the extraction of localized frequency content, which is more suitable for the analysis of frequently varying audio such as music.

Mel-spectrogram is a transformed spectrogram using the mel-scale, a non-linear scale that better approximates the perception of the human auditory system. The approximate formula for the mel-frequency of linear frequency in hertz, f, is as follows:

$$mel(f) = 2595 * \log_{10}\left(1 + \frac{f}{700}\right)$$

To construct a Mel-spectrogram, a parameter for the number of mel-bands, n_{mel}, is chosen (in our models, we used $n_{mel} = 16$). A higher number of mel-bands implies a more detailed representation of the audio data (though too much may lead to overfitting). Using the mel-scale and n_{mel}, a mel-bank filter is constructed and then applied to the audio data to generate the Mel-spectrogram (Figure 1).

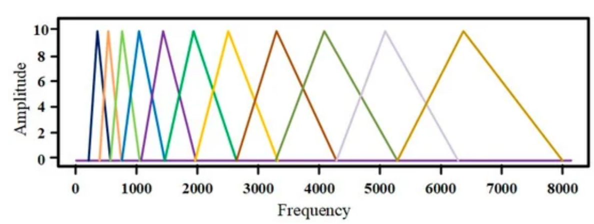

Figure 1. Example mel-bank filter with n_{mel} = 11

An example of a Mel-spectrogram on one of our audio files is shown in Figure 2.

D.3.2 Mel Frequency Cepstral Coefficient (MFCC)

The MFCC is a compact and low-dimension representation of the audio data by applying the Discrete Cosine Transform (DCT) to the Mel-spectrogram. As a result, the most significant cepstral coefficients are extracted. Cepstral coefficients represent the spectral envelope of the audio data and were found to be beneficial as features for machine learning models. To generate the MFCC, a parameter for the number of cepstral coefficients, n_{mfcc}, is chosen (in our models, we used $n_{mfcc} = 13$). An example of an MFCC on one of our audio files is shown in Figure 3.

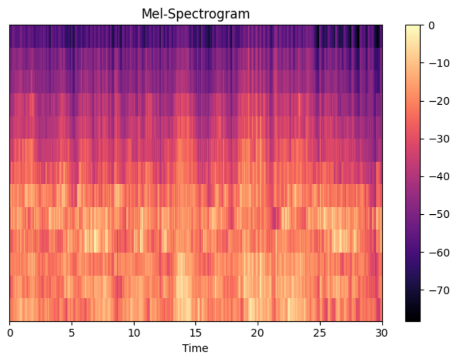

Figure 2. Mel-spectrogram of an audio labeled as classical

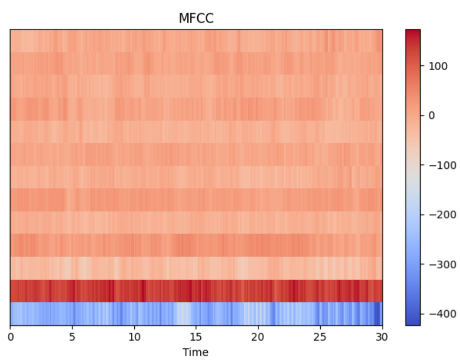

Figure 3. MFCC of the same audio labeled as classical

D.3.3 Chromagram

Chromagram is a type of spectrogram that represents the energy distribution of the 12 standard pitch classes of modern Western music by applying the STFT with respect to the 12 chroma values. Consequently, Chromagrams are invariant to octave differences, timbre, instrumentation, etc.

An example of an Chromagram on one of our audio files is shown in Figure 4.

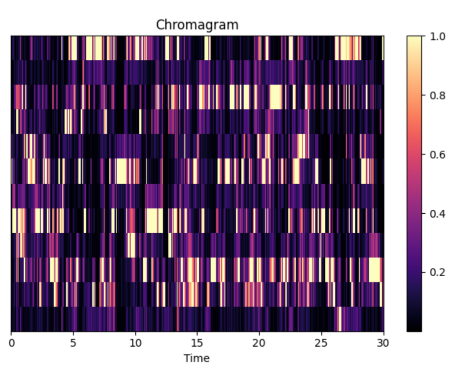

Figure 4. Chromagram of the same audio labeled as classical

D.4. Method

D.4.1 Preprocessing

To prepare the data for training, we performed data augmentation on the dataset. First, each 30-second audio file aforementioned was further split into 5 segments of 6 seconds each. This augmentation increases the amount of labeled data by a factor of 5. Six-second audio files are long enough such that the overarching context (the genre in our case) is maintained while short enough to allow for a higher prediction rate. As a result, there will be 5000 independent audio data each having a genre label and 6 seconds in length.

After the segmentation, audio-visual features are extracted from the audio data. Each 6-second audio file will generate a Mel-spectrogram, MFCC, and Chromagram. These audio-visual features, along with its genre label will be stored in a JSON file to be used later. When we use multiple of these extracted features to represent an audio input, we combined (through concatenation) those features before inputting them into the model. This process of extracting audio-visual features from input audio is illustrated in Figure 5.

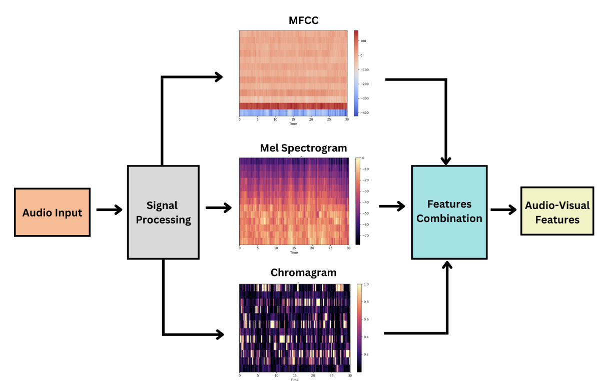

Figure 5. Audio input data processing diagram

D.4.2 Model

We investigated the performance of our model based on 5 different model architectures and 5 different sets of audio-visual features. Specifically, the 5 architectures are CNN, LSTM, Bi-GRU, CNN + LSTM, and CNN + Bi-GRU. The 5 audio-visual features are MFCC, Mel-spectrogram, Chromagram, MFCC + Mel-spectrogram, and MFCC + Mel-spectrogram + Chromagram. We trained and analyzed the performance of each combination of model architecture and audio-visual features, leading to 25 different models that were tested in total.

Below in Figures 6 7 8, we display the architectures of CNN, CNN + LSTM, and CNN + Bi-GRU respectively (the architectures for our Bi-GRU and LSTM models individually are not displayed as they use the same parameters as in the hybrid architectures).

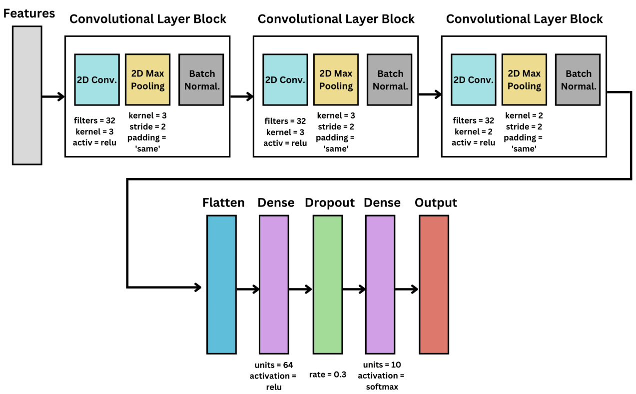

Figure 6. Proposed CNN architecture

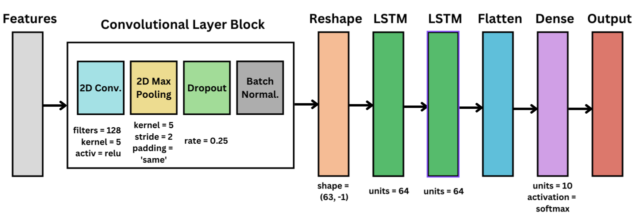

Figure 7. Proposed CNN + LSTM hybrid architecture

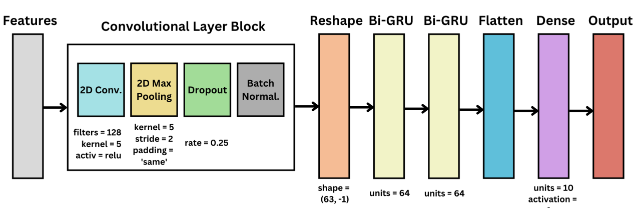

Figure 8. Proposed CNN + Bi-GRU hybrid architecture

D.4.3 Training

To prepare for model training, the audio-visual features JSON file is loaded and then partitioned into a training set (60%), validation set (15%), and testing set (25%). The labels are then one-hot encoded into 10 perpendicular vectors. In our models, we used the Adam optimizer with a learning rate of 0.0001 as an optimizer and categorical cross-entropy for the loss function. As hyperparameters, each model used a batch size of 32 and trained for 30 epochs. In our largest model (CNN + Bi-GRU trained on MFCC + Mel + Chroma features), each epoch to approximately 230 seconds to complete.

D.5. Performance Metrics

The performance of our model is measured by the cumulative test accuracy of the ground truth genre label of the audio data versus the predicted genre label of the same audio data. In other words, the metric is the accuracy of the model, namely the proportion of correctly classified audio data.

E. Experimental Results and Discussion

The summary of the accuracy in our study is presented in Table 1. Moreover, the graph that compares the performance of different audio-visual features on the best-performing architecture, CNN + Bi-GRU hybrid, is shown in Figure 9.

Audio-Visual Features

Model Architecture	MFCC	Mel	Chroma	MFCC + Mel	MFCC + Mel + Chroma
CNN	81.57%	77.70%	55.00%	**84.78%**	82.47%
LSTM	**81.49%**	66.91%	40.17%	74.10%	75.45%
Bi-GRU	**80.34%**	75.22%	38.71%	79.83%	80.28%
CNN + LSTM	84.72%	81.97%	62.81%	**87.30%**	86.40%
CNN + Bi-GRU	89.89%	84.33%	68.37%	88.54%	**89.94%**

Table 1. Summary of the accuracy. For each architecture (row), the highest accuracy achieved is bolded. The overall highest accuracy is in blue.

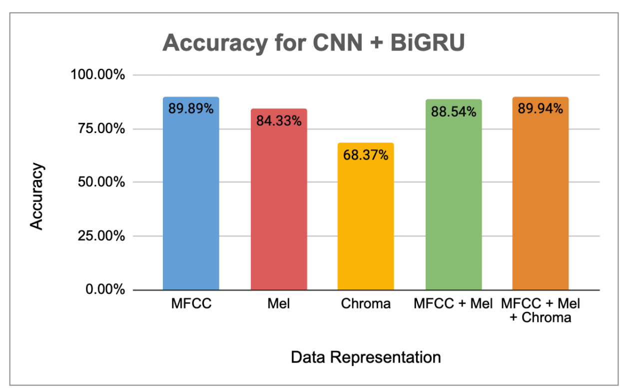

Figure 9. Accuracy with CNN + Bi-GRU hybrid Architecture

The findings indicate that having audio represented in multiple audio-visual features generally leads to better performance. This can be explained by the fact that with more audio features, the model has more characteristics of each genre of music to learn from. There were some architectures, however, where this causal relationship between more features and better performance was not observed. For example, under LSTM architecture, the model achieved 81.49% accuracy with just the MFCC feature, higher than 74.10% (MFCC + Mel) or 75.45% (MFCC + Mel + Chroma) achieved with multiple features. Although we would like to explore this phenomenon more in-depth in the future, we hypothesize that extra features that LSTM fails to learn the relevance of have led to an over-fitted model.

In terms of model architecture, we see that the hybrid of CNN and Bi-GRU layers perform the best in every feature representation. This hybrid that uses both CNN and RNN layers perform better than using each of them alone for it extracts both spatial and temporal information from the audio-visual features. Specifically, with the feature representations of MFCC, Mel, and Chroma, the model achieved 89.94% accuracy in the test set. The plot for Accuracy and Error against Epoch in Figure 10 does not show signs of overfitting since the testing accuracy does not deviate too much from the training accuracy with the increasing number of epochs. This means that the weights generalize well to the dataset. This accuracy is higher than the accuracy achieved by other published models that conduct MCG on the GTZAN dataset using audio-visual features only. The comparison of our model's performance and other state-of-the-art models is shown in Table 2.

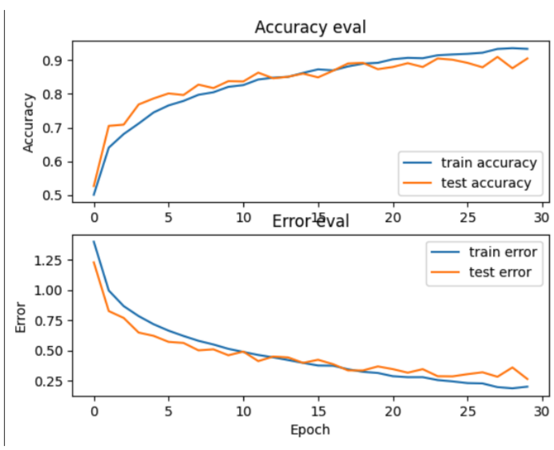

Figure 10. Accuracy and Error of CNN + Bi-GRU hybrid architecture with MFCC + Mel + Chroma features

F. Conclusion and Future Work

In this paper, we found that combining audio-visual features for training was beneficial for musical genre classification, which aligns with our hypothesis. More specifically, our CNN + Bi-GRU architecture achieved an accuracy of 89.94% using a combination of MFCC + Mel + Chroma features on the GTZAN dataset. Our model outperformed

Table 2. Comparison of State-of-the-Art Models

Method	Accuracy
George Tzanetakis [13]	61.00%
G. Sun et al. [12]	66.40%
A Heaki et al. [7]	70.60%
Nilesh M. [10]	77.78%
Praseneet Fulzeele et al. [6]	89.00%
N. Farajzadeh [5]	86.00%
Pradeep Kumar D. et al. [9]	86.00%
Jan Jakubik [8]	87.70%
Mohsin Ashraf et al. [1]	89.30%
Our Model	**89.94%**

other state-of-the-art models that only used audio-visual features such as the 89.30% accuracy achieved in January 2023 [1].

Due to time and computing constraints, we understand there are many potential improvements that could be made to our model. For future works, we will incorporate other audio features such as spectral centroid, tempo extraction, percussive features, harmonic feature, etc. as presented in Jinliang L. et al 2021 [2], which achieved an accuracy of 93.4% by using both Visual Feature Extraction module and Audio Feature Extraction module. One approach is to utilize the statistics provided in the GTZAN dataset. We will also try higher resolution of audio-visual feature extraction by increasing the number of mel-bands for generating Mel-spectrograms (this was again due to limited computational power). Additionally, we will tweak and optimize hyperparameters (i.e. number of layers, epoch, etc.), train on other audio datasets (such as the Million Song Dataset), and investigate other ML architecture models (e.g. transformers).

G. Individual Contributions

- Literature reviews (Joey Zheng and Daniel Kim)
- Audio data download & feature extraction & preprocessing (Joey Zheng)
- Model layer architecture design & implementation (Daniel Kim)
- Training and testing of different models (Joey Zheng and Daniel Kim)
- Final report & presentation (Joey Zheng and Daniel Kim)

References

[1] M. Ashraf, F. Abid, I.U. Din, J. Rasheed, M. Yesiltepe, S.F. Yeo, and M.T. Ersoy. A hybrid cnn and rnn variant model for music classification. *Applied Sciences*, 13, 2023. 1, 5

[2] Jia Dai, Wen-Ju Liu, Chongjia Ni, Like Dong, and Hong Yang. "multilingual" deep neural network for music genre classification. 09 2015. 1, 5

[3] Tyler Dammann and Kevin Haugh. Genre classification of spotify songs using lyrics, audio previews, and album artwork, 2017. 1

[4] J Stephen Downie. Music information retrieval. *Annual review of information science and technology*, 37(1):295–340, 2003. 1

[5] N. Farajzadeh, N. Sadeghzadeh, and M. Hashemzadeh. Pmg-net: Persian music genre classification using deep neural networks. *Entertainment Computing*, 44:100518, 2023. 5

[6] P. Fulzele, R. Singh, N. Kaushik, and K. Pandey. A hybrid model for music genre classification using lstm and svm. In *Proceedings of the 2018 Eleventh International Conference on Contemporary Computing (IC3)*, pages 1–3. IEEE, 2018. 5

[7] A. Heakl, A. Abdelgawad, and V. Parque. A study on broadcast networks for music genre classification. In *Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, Padua, Italy, July 2022. 5

[8] J. Jakubik. Evaluation of gated recurrent neural networks in music classification tasks. In *Proceedings of the 38th International Conference on Information Systems Architecture and Technology—ISAT 2017*, pages 27–37. Springer, 2017. 5

[9] D.P. Kumar, B.J. Sowmya, Chetan, and K.G. Srinivasa. A comparative study of classifiers for music genre classification based on feature extractors. In *2016 IEEE Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*, pages 190–194. IEEE, 2016. 5

[10] N. M. Patil and M. U. Nemade. Music genre classification using mfcc, k-nn and svm classifier. *Int. J. Comput. Eng. Res. Trends*, 4:2349–7084, 2017. 5

[11] Garima Sharma, Kartikeyan Umapathy, and Sridhar Krishnan. Trends in audio signal feature extraction methods. *Applied Acoustics*, 158:107020, 2020. 1

[12] G. Sun. Research on architecture for long-tailed genre computer intelligent classification with music information retrieval and deep learning. *Journal of Physics: Conference Series*, 2033:012008, 2021. 5

[13] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10:293–302, 2002. 1, 5