RhythmiQ: A Deep Learning-Based Music Genre Classifier

Anshit Agarwal (b23cs1087@iitj.ac.in)
Kaustubh Salodkar (b23ee1033@iitj.ac.in)
Supervised By : Pratik Mazumder

Indian Institute of Technology, Jodhpur

April 17, 2025

Abstract

Music genre classification is a fundamental task in music information retrieval systems. This project, named RhythmiQ, explores multiple machine learning approaches to develop an accurate classification system for identifying music genres using the GTZAN dataset. Five models have been implemented and evaluated: (1) Support Vector Machine (SVM), (2) Long Short-Term Memory (LSTM), (3) LSTM + SVM Hybrid, (4) Convolutional Neural Network (CNN), and (5) Residual-Gated CNN (Res-Gated CNN). Our experimental results demonstrate that CNN-based architectures achieve the highest accuracy (86.44%), significantly outperforming traditional methods. Future work focuses on implementing transformer-based models, improving the Res-Gated CNN architecture, and developing a real-time genre prediction application.
Keywords: Mel Spectrograms, GTZAN Dataset, Audio Signal Processing

Contents

1 Introduction

Music genre classification is a challenging problem in audio signal processing and machine learning. The RhythmiQ project evaluates five different machine learning and deep learning models using the GTZAN dataset, containing 1,000 audio tracks across 10 music genres. Our aim is to develop an accurate and efficient recognition system for classifying music genres. Automatic genre classification offers

numerous practical applications for music streaming services, producers, libraries, and researchers. It enhances recommendation systems, assists in music analysis, enables efficient cataloging of collections, and supports quantitative research of genre-specific patterns. The RhythmiQ system consists of four

key components: audio preprocessing, feature extraction, model training, and performance evaluation. Through a combination of carefully engineered features and advanced deep learning architectures, we address the challenges of subjective genre boundaries and complex audio signals.

Table 1 provides a sample of the extracted audio features from the GTZAN dataset's blues genre recordings, illustrating the rich feature set used in our classification models.

Table 1: Sample of Extracted Audio Features from GTZAN Dataset (Blues Genre)

filename	**length**	**chroma_stft_mean**	**rms_mean**	**zero_crossing_rate**	**tempo**
blues.00000.0.wav	66149	0.335406	0.130405	0.081851	129.199
blues.00000.1.wav	66149	0.343065	0.112699	0.087173	123.047
blues.00000.2.wav	66149	0.346815	0.132003	0.071383	123.047
blues.00000.3.wav	66149	0.363639	0.132565	0.069426	123.047

2 Dataset

We utilize the GTZAN Genre Collection, which is a standard benchmark dataset for music genre classification tasks. The dataset consists of 10 genres arranged in the following table:

Table 2: GTZAN Dataset Genre Categories

Music Genres				
Blues	Classical	Country	Disco	Hip hop
Jazz	Metal	Pop	Reggae	Rock

Each original audio file in the dataset is 30 seconds long. To enhance data diversity and improve learning, we segment each audio file into 3-second clips. After preprocessing, the dataset contains:

- 10 genres
- 1000 samples per genre
- Total: 10,000 audio clips

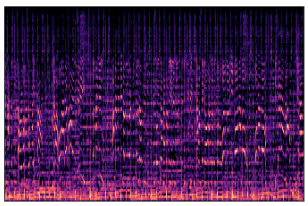


Figure 1: Disco

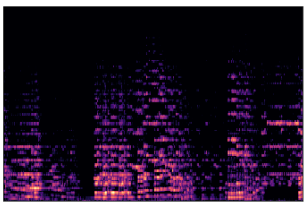


Figure 2: Classical

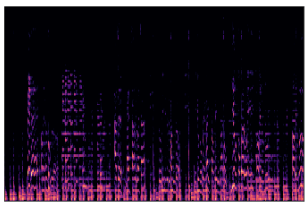


Figure 3: Blues

3 Feature Extraction

For our classification task, we utilized pre-extracted features from the `features_3_sec.csv` file provided with the GTZAN dataset. This file contains 55 audio features extracted from each 3-second clip, categorized as follows:

Table 3: Extracted Audio Features

Category	**Feature Names**
Chroma STFT	`chroma_stft_mean`, `chroma_stft_var`
Spectral Centroid	`spectral_centroid_mean`, `spectral_centroid_var`
Spectral Bandwidth	`spectral_bandwidth_mean`, `spectral_bandwidth_var`
Spectral Rolloff	`rolloff_mean`, `rolloff_var`
Zero Crossing Rate	`zero_crossing_rate_mean`, `zero_crossing_rate_var`
Harmonic Content	`harmony_mean`, `harmony_var`
Perceptual Spread	`perceptr_mean`, `perceptr_var`
Tempo	`tempo`
MFCC Means	`mfcc_1_mean` to `mfcc_20_mean`
MFCC Variances	`mfcc_1_var` to `mfcc_20_var`

Additionally, we generate Mel Spectrograms from the audio clips to serve as input for our deep learning models, particularly the CNN and Res-Gated CNN architectures.

4 Models Implemented

4.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a traditional machine learning algorithm that constructs a hyperplane in high-dimensional space to separate different classes. For our implementation, we utilized the 55 extracted audio features as input to the SVM model. We experimented with different kernel functions, including linear, polynomial, and radial basis function (RBF), ultimately selecting the RBF kernel due to its superior performance on our dataset.

4.2 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) architecture designed to model temporal sequences and their long-range dependencies. Our LSTM model processes the sequential nature of music by analyzing features across time frames. The architecture is detailed below:

Table 4: LSTM Architecture for Music Genre Classification

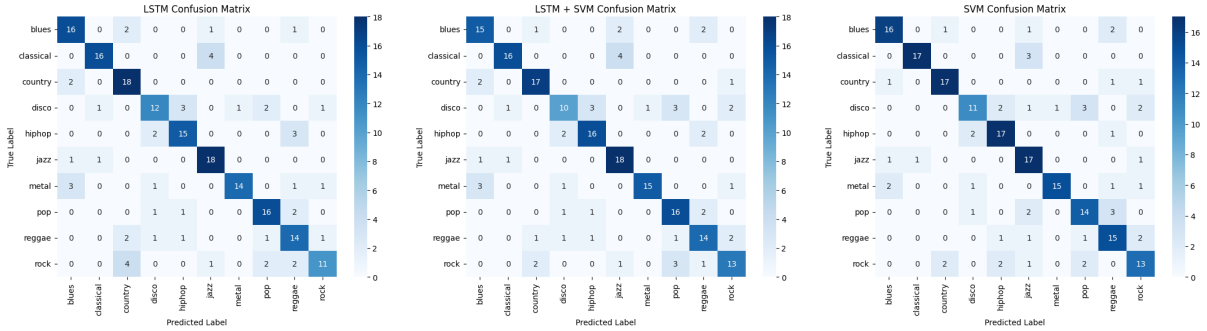
Layer (type)	**Output Shape**	**Param #**
lstm (LSTM)	(None, 1, 128)	94,208
dropout (Dropout)	(None, 1, 128)	0
lstm_1 (LSTM)	(None, 64)	49,408
dropout_1 (Dropout)	(None, 64)	0
dense (Dense)	(None, 32)	2,080
dense_1 (Dense)	(None, 10)	330

This model leverages two stacked LSTM layers with dropout (0.3) for regularization, followed by dense layers for classification. The first LSTM layer returns sequences which are fed into the second LSTM layer, producing a single vector output that captures temporal patterns across the entire audio segment. Despite its sophisticated sequential learning capability, the LSTM model achieved 75% accuracy, suggesting that for short music clips, spatial features may be more discriminative than temporal patterns.

4.3 LSTM + SVM Hybrid

The LSTM + SVM hybrid model combines the sequential learning capabilities of LSTM with the classification strength of SVM. In this approach, the LSTM network serves as a feature extractor, with its hidden layer outputs fed into an SVM classifier for final genre prediction. This hybrid approach aims to leverage the strengths of both models:

- LSTM captures temporal patterns in music
- SVM provides robust classification boundaries



4.4 Convolutional Neural Network (CNN)

Our CNN model processes the extracted audio features in a 2D format, treating them as image-like data. The model architecture is summarized below:

Table 5: CNN Architecture for Music Genre Classification

Layer (type)	**Output Shape**	**Param #**
conv2d_17 (Conv2D)	(None, 18, 1, 32)	128
max_pooling2d_17 (MaxPooling2D)	(None, 9, 1, 32)	0
dropout_14 (Dropout)	(None, 9, 1, 32)	0
flatten_9 (Flatten)	(None, 288)	0
dense_18 (Dense)	(None, 128)	36,992
dropout_15 (Dropout)	(None, 128)	0
dense_19 (Dense)	(None, 10)	1,290

This lightweight CNN architecture processes the 55 audio features through convolutional, pooling, and dense layers with strategic dropout (0.3) to prevent overfitting. The model effectively learns discriminative patterns from the audio features, resulting in the highest accuracy 86.44% among all implemented models.

4.5 Residual-Gated CNN (Res-Gated CNN)

Our Res-Gated CNN enhances the standard CNN architecture by incorporating residual connections and gating mechanisms to control information flow and address the vanishing gradient problem. The detailed architecture is shown below:

This complex architecture features multiple residual blocks, each consisting of convolutional layers followed by gating mechanisms implemented via Lambda layers. The model includes both global average and max pooling operations, followed by dense layers with dropout for regularization. While achieving a strong 84.18% accuracy, this architecture presents opportunities for further optimization through hyperparameter tuning.

Table 6: Residual-Gated CNN Architecture for Music Genre Classification

Layer (type)	**Output Shape**	**Param #**	**Connected to**
input_layer (InputLayer)	(None, 55, 1)	0	-
conv1d (Conv1D)	(None, 55, 64)	256	input_layer[0][0]
conv1d_1 (Conv1D)	(None, 55, 128)	24,704	conv1d[0][0]
lambda_1 (Lambda)	(None, 55, 64)	0	conv1d_1[0][0]
lambda (Lambda)	(None, 55, 64)	0	conv1d_1[0][0]
activation (Activation)	(None, 55, 64)	0	lambda_1[0][0]
multiply (Multiply)	(None, 55, 64)	0	lambda[0][0], activation[0][0]
add (Add)	(None, 55, 64)	0	multiply[0][0], conv1d[0][0]
max_pooling1d (MaxPooling1D)	(None, 27, 64)	0	add[0][0]
... Additional residual blocks with similar structure ...			
global_average_pooling1d	(None, 128)	0	max_pooling1d_3[0][0]
global_max_pooling1d	(None, 128)	0	max_pooling1d_3[0][0]
concatenate (Concatenate)	(None, 256)	0	global_avg/max_pooling1d[0][0]
dense (Dense)	(None, 200)	51,400	concatenate[0][0]
dropout (Dropout)	(None, 200)	0	dense[0][0]
dense_1 (Dense)	(None, 100)	20,100	dropout[0][0]
dropout_1 (Dropout)	(None, 100)	0	dense_1[0][0]
dense_2 (Dense)	(None, 10)	1,010	dropout_1[0][0]

5 Experiments and Results

We trained and evaluated each model using the same train-test split to ensure a fair comparison. The dataset was divided into 80% training and 20% testing sets.

Table 7: Classification Accuracies on Test Dataset

Model	**Test Accuracy (%)**
Convolutional Neural Network (CNN)	86.44
Residual-Gated CNN (Res-Gated CNN)	84.18
Support Vector Machine (SVM)	77.00
Long Short-Term Memory (LSTM)	75.00
LSTM + SVM Hybrid	75.00

The results demonstrate that CNN-based architectures significantly outperform traditional methods for music genre classification. The standard CNN achieved the highest accuracy at 86.44%, closely followed by the Res-Gated CNN at 84.18%. SVM performed reasonably well with 77% accuracy, while LSTM-based models achieved 75% accuracy.

6 Discussion

6.1 Performance Analysis

The superior performance of CNN-based models can be attributed to their ability to capture spatial patterns in mel spectrograms effectively. These patterns correspond to genre-specific characteristics such as rhythmic structures, harmonic content, and timbral features. The slightly lower performance of the Res-Gated CNN compared to the standard CNN suggests potential areas for improvement, such as hyperparameter tuning or addressing training instability.

SVM demonstrated strong performance as a traditional machine learning baseline, indicating that the handcrafted features contain significant discriminative information for genre classification. The relatively lower performance of sequence models like LSTM suggests that for short 3-second clips, spatial patterns may be more informative than temporal dependencies.

6.2 Genre-wise Performance

Analysis of genre-wise classification performance revealed that certain genres were more easily distinguishable than others:

- Classical and metal music were the most accurately classified genres across all models, likely due to their distinctive spectral characteristics.
- Rock and country music showed the highest confusion rates, particularly with each other, reflecting their similar instrumentation and structural elements.
- Hip hop and disco showed moderate confusion rates with other genres.

7 Future Work

Based on our current results, we have identified several directions for future work:

- Implementing transformer-based models to capture long-range dependencies in audio data
- Improving the Res-Gated CNN architecture through hyperparameter optimization and architectural refinements
- Exploring ensemble methods combining predictions from multiple models
- Developing a real-time genre prediction application using Streamlit
- Expanding the dataset to include more diverse music styles and newer genres

8 Conclusion

In this project, we developed and evaluated five different models for music genre classification using the GTZAN dataset. Our results demonstrate that CNN-based architectures achieve the highest accuracy, with the standard CNN model reaching 86.44% test accuracy. These findings highlight the effectiveness of deep learning approaches, particularly convolutional networks, for audio classification tasks. The

project establishes a strong foundation for music genre classification and provides insights into the relative performance of different machine learning paradigms. The planned improvements, especially the implementation of transformer-based models and the development of a real-time application, represent valuable next steps to enhance both accuracy and practical utility of our music genre classification system.

A Contribution of Each Member

1. Anshit Agarwal (b23cs1087@iitj.ac.in): Implemented CNN and Res-Gated CNN models, project architecture and feature extraction.
2. Kaustubh Salodkar (b23ee1033@iitj.ac.in): Implemented SVM, LSTM, and SVM+LSTM hybrid models, created visualizations, and developed evaluation metrics.