Harmonizing Genres: An Ensemble Approach Using CNNs and RNNs for Music Genre Classification

CS342 Neural Networks Course Project

Jatin Kulkarni
Cornell Tech
jatinkulkarni.com
Work completed at The University of Texas at Austin
jkulkarnics@utexas.edu

Abstract

Music genre classification remains a difficult problem due to the diversity, overlap, and subjectivity of musical styles. In this project, we explore an ensemble-based approach that combines Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to capture both spectral and temporal characteristics of music audio. Starting with the GTZAN dataset and augmenting it with the Hugging Face `ccmusic-database/music_genre` dataset, we investigate multiple audio representations including raw waveforms, spectrograms, log-mel spectrograms, and chromagrams. We first train specialized CNN models for each representation, then extend the approach with LSTM- and GRU-based recurrent models to better capture sequential structure such as melody, harmony, and rhythm. While classification accuracy remained modest, the project provided useful insight into representation choice, architectural trade-offs, and the complexity of genre prediction across heterogeneous audio inputs.

1 Introduction

Music genre classification is an important task in recommendation systems, music discovery, playlist generation, and large-scale audio indexing. However, accurate genre prediction is challenging because genres are often defined by a mixture of instrumentation, texture, rhythm, melody, production style, and cultural convention rather than by a single objective signal. Many genres also share overlapping acoustic characteristics, which makes it difficult for machine learning models to separate them reliably.

This project was motivated by the idea that different model families may capture different aspects of music. Convolutional Neural Networks (CNNs) are effective at learning local spatial patterns in image-like representations such as spectrograms, while Recurrent Neural Networks (RNNs) are better suited for modeling temporal dependencies and sequential structure. Our goal was to investigate whether a combined approach could improve genre prediction by leveraging the complementary strengths of both architectures.

More specifically, the project aimed to use ensemble learning to integrate CNN-based image-style audio analysis with RNN-based sequence modeling. We hypothesized that CNNs would perform well at extracting local spectral patterns, while RNNs would better capture the temporal progression of rhythm, melody, and harmonic movement that often distinguishes related genres.

2 Related Motivation

Recent work in machine learning for music has shown that no single representation fully captures the richness of audio. Raw waveforms preserve the original signal but are difficult to model directly. Spectrogram-based representations expose time-frequency structure and are often more amenable to deep learning. Chroma-based features emphasize pitch class information and can better expose harmonic content. These observations motivated us to build multiple representation-specific models rather than relying on a single preprocessing pipeline.

Our project also took inspiration from ensemble learning, where predictions from multiple models are combined to improve robustness. In a task such as music genre classification, ensemble methods are especially appealing because different genres may be separable along different dimensions: timbre, harmonic structure, rhythmic repetition, or temporal evolution.

3 Dataset and Preprocessing

We constructed our dataset using the GTZAN benchmark dataset together with additional samples from the Hugging Face `ccmusic-database/music_genre` dataset. Audio files were standardized into 30-second snippets to ensure a consistent input length across models. This allowed both CNN and RNN architectures to operate on fixed-size representations while still preserving enough musical context to capture genre-defining structure.

To challenge the models and improve robustness, we introduced controlled variability into the data. In addition to standard audio clips, we included manipulated and alternative forms of songs such as instrumental tracks, acapella tracks, and sped-up versions. The goal was to evaluate whether models could generalize beyond conventional full-mix recordings and still recognize genre-related features under altered acoustic conditions.

After segmentation, each clip was transformed into one or

more feature representations:

- **Waveforms**: raw audio amplitude sequences
- **Spectrograms**: time-frequency magnitude representations
- **Log-Mel Spectrograms**: perceptually weighted frequency representations
- **Chromagrams**: pitch-class intensity distributions over time

These representations were selected because they emphasize different musical properties. For example, chromagrams highlight harmonic progression, while log-mel spectrograms better reflect the perceptual structure of sound.

4 Methodology

4.1 CNN Models

The first stage of the project focused on CNN-based models trained independently on different audio representations. For image-like inputs such as spectrograms, log-mel spectrograms, and chromagrams, we used 2D convolutional architectures to extract local time-frequency patterns. For raw waveforms, we used 1D convolutions to learn temporal amplitude features directly from the signal.

These models were trained using standard optimization methods such as stochastic gradient descent (SGD) and Adam. Dropout regularization was incorporated to reduce overfitting and improve generalization. By building separate CNNs for each representation, we aimed to identify which forms of preprocessing were most useful for genre discrimination.

4.2 RNN Models

Although the CNN models captured useful local patterns, they were limited in their ability to model the sequential nature of music over time. To address this, we incorporated recurrent architectures, primarily LSTMs and GRUs, to better capture temporal dependencies such as rhythmic evolution, melodic contour, and harmonic progression.

We designed multiple RNN variants tailored to specific feature types:

Chroma RNN. The chroma-based RNN modeled the evolution of the 12 pitch classes through time. This representation was particularly useful for understanding harmonic progression and tonal movement, which can be highly informative for genre identification.

Log-Mel Spectrogram RNN. The log-mel RNN processed sequences of perceptually relevant spectral features. This helped model textural differences and timbral evolution, especially for genres that may have similar tempos but different sound profiles.

Spectrogram RNN. The spectrogram-based RNN attempted to learn temporal changes in the frequency spectrum directly. This representation was intended to capture patterns such as recurring rhythmic emphasis, dense instrumental layering, or sustained melodic structure.

Waveform RNN. Finally, we explored a waveform-based recurrent model that consumed raw temporal signal data. While more difficult to optimize, this representation preserved fine-grained temporal information related to beat, rhythm, and transient structure.

4.3 Ensemble Learning

To combine the strengths of the individual models, we used an ensemble strategy over the CNN and RNN outputs. We explored straightforward methods such as weighted averaging and simple stacking. The intuition was that different models would capture complementary information, and combining them could mitigate the weaknesses of any single representation or architecture.

The ensemble was intended not only to improve raw accuracy, but also to increase robustness across different kinds of songs and transformed audio variants.

5 Experiments and Results

Our experiments focused on genre prediction across a subset of genres including Pop, Rock, Classical/Opera, and R&B. We evaluated the ability of both standalone models and the ensemble to classify standard clips as well as altered inputs such as instrumental, acapella, and sped-up versions.

The CNN models achieved approximately $\sim$ 26–28% classification accuracy across the tested genres. These results suggested that image-like audio representations were informative but insufficient on their own for reliable genre prediction. In particular, genres with overlapping timbral or rhythmic patterns were difficult to separate using only local spectral features.

The addition of RNNs improved the project conceptually by providing a mechanism for modeling sequential dependencies, especially in tasks where melody and progression mattered more than instantaneous texture. However, the overall system still struggled to exceed modest accuracy levels. The results reinforced the idea that genre classification is inherently difficult and that representation and dataset limitations can dominate model performance.

6 Discussion

Several key observations emerged from the project.

First, **model architecture matters**. CNNs and RNNs did not fail in the same way: CNNs were effective at extracting local patterns from structured inputs, while RNNs were better aligned with the temporal nature of musical sequences.

Second, **feature representation matters**. Different input representations surfaced different musical attributes. Chromagrams emphasized harmonic content, log-mel spectrograms captured perceptually relevant timbre, and waveforms preserved raw timing information. No single representation was sufficient on its own.

Third, **genre boundaries are often ambiguous**. Many genres share notes, frequency patterns, instrumentation, and production characteristics. This overlap makes genre prediction substantially harder than tasks with cleaner class separation.

Fourth, **data imbalance affected performance**. Some genres had fewer high-quality examples than others, which reduced the model's ability to generalize evenly across classes. Smaller subsets, such as opera or classical-oriented examples, were particularly vulnerable to underrepresentation.

Overall, the project highlighted that music genre classification is not just a modeling problem; it is also a representation, labeling, and dataset construction problem.

7 Future Work

There are several directions that could improve upon this work.

One natural extension would be to use transformer-based audio architectures, which can model longer-range dependencies more effectively than standard recurrent models. Another would be to leverage pretrained audio encoders or representation learning methods rather than training all models from scratch. More advanced augmentation techniques and better-balanced datasets could also improve robustness and reduce bias toward dominant genres.

Finally, a more detailed evaluation framework could help determine whether the models are learning genre semantics or simply exploiting superficial production cues.

8 Conclusion

This project explored a hybrid and ensemble-based approach to music genre classification using CNNs and RNNs across multiple audio representations. While overall performance remained limited, the work provided a useful comparative study of how representation choice and model architecture affect genre prediction. The project also underscored the difficulty of modeling musical style, particularly when class boundaries are subjective and acoustically overlapping. Even with modest results, the study offered meaningful insight into the strengths and limitations of mixed-model approaches for music analysis.

Acknowledgments

This work was completed at The University of Texas at Austin as part of the CS342 Neural Networks course in Spring 2024.

Notes

This document is an adapted portfolio version of an undergraduate course project. Due to university account access restrictions, the original code, notebook artifacts, and some intermediate visualizations are no longer available. The paper text has been reconstructed and consolidated from saved project materials.