Deep Learning For Enhanced Music Classification and Recommendation

Zane Perry

CSCI 5922, University of Colorado Boulder

Abstract. This work presents a deep learning approach for music classification and recommendation that integrates digital signal processing (DSP) techniques with neural network architectures. By converting raw audio signals into structured transform representations, the system captures both spectral and temporal characteristics of music. These features are used as input to a hybrid model combining Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), enabling nuanced modeling of musical structure and user preferences. The system is evaluated using a custom dataset collected via user surveys, with models trained to predict user-song compatibility through binary classification. Results indicate that single-modal CNN architectures achieve competitive accuracy with significantly lower computational cost than multimodal alternatives. These findings highlight the potential of acoustically-driven recommendation systems and point to future improvements with larger and more diverse datasets.

1 Introduction

Music plays a central role in entertainment, culture, and personal expression, making effective music recommendation and classification valuable for enhancing user experience in streaming services and music media. Music recommendation systems heavily impact listener satisfaction, music discovery, and content personalization. Improving these systems through advanced signal processing and deep learning techniques could potentially provide improvements in recommendation accuracy, music discovery, and user satisfaction.

Existing music classification and recommendation solutions often rely on metadata or simple acoustic features, which fail to fully capture the nuanced characteristics and emotional content within music. This limitation leads to generic recommendations and a disconnect from the listener's unique tastes. Furthermore, traditional methods tend to overlook the actual musical structure in raw audio data, resulting in formulaic interpretations of musical taste rather than actual musical content.

To overcome these limitations, I propose integrating digital signal processing (DSP) techniques with advanced deep neural network architectures to achieve robust and accurate music classification and recommendation. My approach leverages spectrogram analysis and temporal feature extraction combined with convolutional and recurrent neural networks to create an architecture that captures deeper insights into musical structure and listener preferences.

2 Related Work

A large body of research has explored the intersection of digital signal processing (DSP), deep learning, and music information retrieval. Foundational work by Müller et al. (2011) established core DSP techniques for music analysis, such as beat tracking, pitch extraction, and source separation, emphasizing how spectral representations like spectrograms and chroma features can capture perceptual aspects of audio signals [1]. More recently, Mycka and Mańdziuk (2024) reviewed emerging trends in applying artificial intelligence to music, identifying a critical challenge: bridging the gap between low-level signal features and high-level musical understanding [2]. While these studies demonstrate the power of DSP for analyzing musical content, they generally focus on standalone analysis rather than using these techniques as input pipelines to neural networks. In contrast, my work integrates DSP directly into a feature extraction pipeline optimized for deep learning, transforming representations such as Mel-Frequency Cepstral Coefficients (MFCCs), Short-Time Fourier Transforms (STFT), and Discrete Wavelet Transforms (DWT) into structured inputs for a hybrid CNN–RNN architecture.

In the domain of music classification, various approaches have leveraged convolutional neural networks to learn musical patterns from spectral representations. Zhang and Li (2025) proposed a parallel CNN architecture for genre classification, enhancing performance using an optimization algorithm to refine network weights [3]. Elbir and Aydin (2020) explored genre classification and recommendation jointly, also relying on CNNs but without explicit modeling of temporal progression [4]. My architecture builds on these efforts by combining CNNs with recurrent layers—specifically Long Short-Term Memory (LSTM) networks—to capture both spectral structure and temporal evolution. This hybrid architecture is particularly effective in modeling how music unfolds over time, going beyond simple static classification of genre or style.

Music recommendation has also seen increasing use of deep learning, with an emphasis on user modeling and similarity metrics. Schedl (2019) reviewed deep learning-based music recommender systems, pointing out the growing use of embeddings and end-to-end learning to replace traditional collaborative filtering or metadata-based approaches [5]. Zhang (2022) developed a CNN-based recommendation model using only audio features, but without incorporating sequence modeling [6]. Damak et al. (2021) proposed a hybrid method using sequence modeling for explainable recommendations, but relied more on user behavior data than acoustic content [7]. My approach distinguishes itself by leveraging DSP-derived acoustic features to drive personalized recommendation directly from the musical signal, rather than relying on user interaction graphs or metadata. This allows for more nuanced modeling of listener preferences based on musical content alone.

Taken together, these works demonstrate the strengths of both DSP and deep learning in music analysis, classification, and recommendation. However, the integration of these domains—using DSP to produce input features that are fed through temporally-aware deep learning models—remains underexplored. My proposed system fills this gap by tightly coupling signal-level feature extraction with CNN-based spatial modeling and RNN-based sequence learning, enabling the system to learn from both the sound characteristics and the temporal structure of music in a unified architecture

3 Methodology

The core of this experiment is designed around a survey of music preferences collected from students around CU Boulder. The survey begins by asking the participant what their favorite song is, which will eventually be used as input for the model. The participant is then asked to review a list of 50 mainstream popular songs that span multiple genres, and mark simply whether they like or dislike the song. Finally, the participant is asked to choose their favorite 3 songs out of the provided list. This survey was designed to extract as much data as possible from those who filled it out while also collecting the most diverse data possible. This led to the selection of using mainstream songs for the survey as providing songs that a majority of people would know without having to listen to it would mean less time would be spent evaluating each song, and more songs could be included in the survey overall. Additionally, the inclusion of multiple different genres provided the diverse audio features desired for the model to learn as much as possible. 75 responses were collected using the survey.

This project begins by converting each song's audio signal into visual representations that describe how its sound evolves over time. These representations include Mel-Frequency Cepstral Coefficients (MFCCs), which capture the timbre or tone quality of a sound, and optionally Short-Time Fourier Transform (STFT) and Discrete Wavelet Transform (DWT), which describe how different frequencies are present at different times. These techniques are all part of digital signal processing (DSP) and allow the computer to understand patterns in sound similarly to how humans perceive melody, rhythm, and texture.

The MFCCs are structured like a spectrogram: a 2D image where one axis represents pitch-related information and the other represents time. Instead of analyzing the whole image at once, the model breaks it into a sequence of short "snapshots" (small time segments). Each snapshot is first passed through a Convolutional Neural Network (CNN) — a type of neural network that is very good at recognizing visual patterns. The CNN identifies meaningful features like harmonic contours or rhythmic textures in each frame.

Once the CNN has processed the entire sequence of frames, the outputs are passed into a second neural network called a Long Short-Term Memory (LSTM) network. The LSTM is a special kind of Recurrent Neural Network (RNN) that is designed to recognize patterns across time. This lets the model understand how a song changes from beginning to end — for example, how a beat drops or how a chorus repeats. This combination of CNN followed by LSTM allows the model to recognize both short-term and long-term patterns in the music.

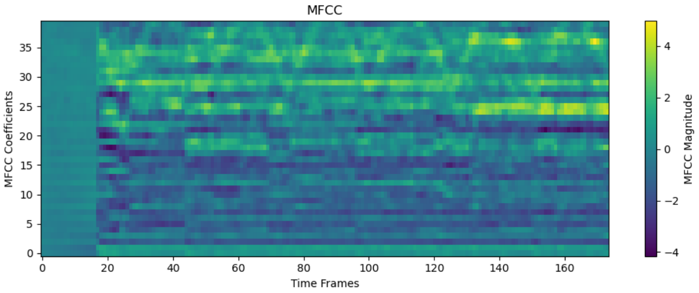
The final output of the model is a set of predictions for each song. These predictions are used either to classify the song into relevant categories or to determine how well the song matches a user's preferences. The model is trained using a binary classification approach, meaning it learns to predict a series of "yes" or "no" answers — for example, "Would this user like this song?" or "Is this song similar to others in its cluster?"

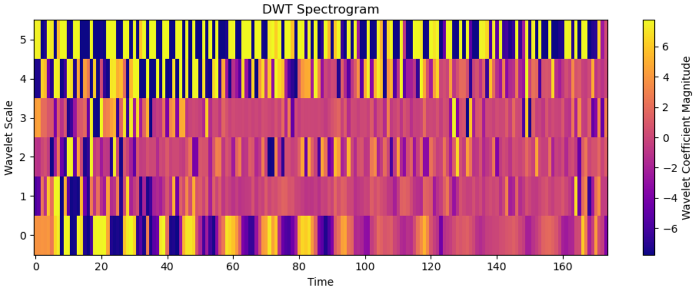
The training process uses standard neural network techniques like dropout (to prevent overfitting) and carefully chosen weight initialization strategies to ensure stable learning. Evaluation is based on Hamming accuracy, which measures how many of the model's predictions were correct across multiple possible answers.

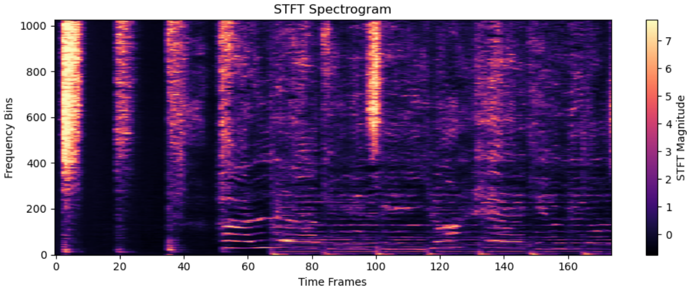
By combining handcrafted audio features with a neural network designed to understand both local details and long-term structure, this approach provides a powerful tool for understanding musical content and recommending songs in a personalized way.

4 Experiments

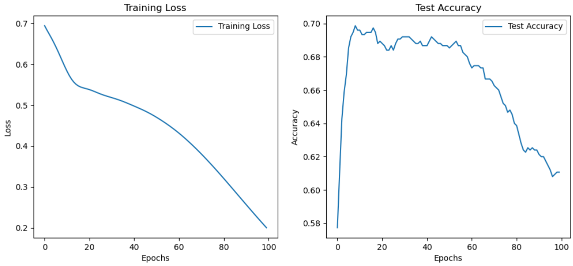
The first step of the research for this project was to test single-modal models that consider each audio feature separately and test their performance on the data individually. After considering different architectures, it was decided that a CNN to RNN hybrid model would be used for the MFCC data, because the output data had both spatial and temporal relationships across the entire audio signal; The CNN would be able to find patterns within small pockets of time, while the RNN would be able to analyze the entire piece of music as a whole. For the DWT and STFT features, because the resulting output was only a single snapshot of the entire audio signal, a CNN was used to find relations in the structure and temporal effects were not considered. Future work on this model could include using sliding time windows of these features in order to consider their evolution through time in the audio. An example output of each of the signal transforms is included below.



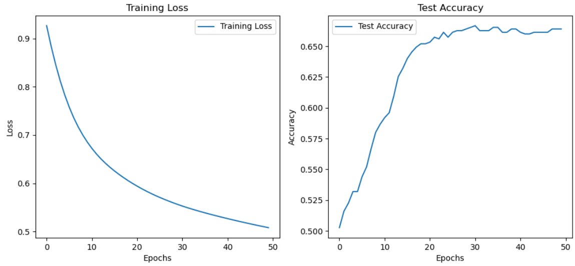


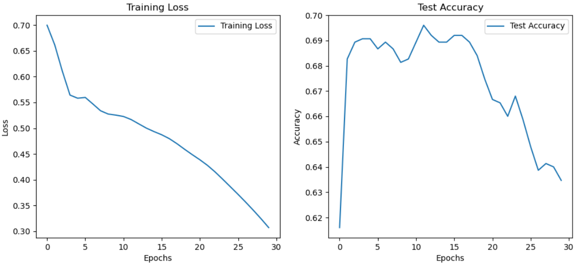


For each model, several hyperparameters such as epochs, dropout, learning rate, and weight initialization were considered until the best performing and most stable configuration was found. The analysis of the DWT model will be used as an example for how each model was designed. The first thing considered once a base model was functioning was how many epochs to train the model for, and what learning rate to use. Training the model for 100 epochs with a learning rate of 0.001 resulted in the following training loss and test accuracy.

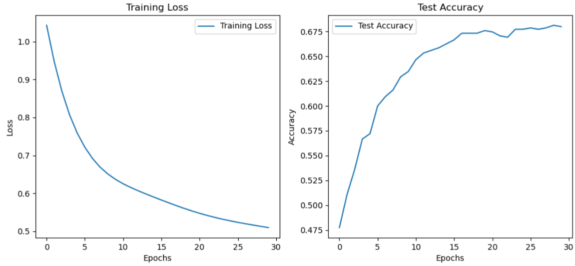


Despite the fact that the training loss continuously decreased, the accuracy initially spiked upwards and decreased over time which is a defining characteristic of overfitting. Decreasing the learning rate to 0.0005 did not lead to a more accurate convergence and just took longer to settle into a single value, and increasing the learning rate to 0.005 led to large oscillations that never really converged to a specific accuracy, implying that the rate was too large to find a good optimal weight configuration.





Because of these results, a standard configuration of 30 training epochs with a learning rate of 0.001 was decided upon. Different weight initializations were also included, and while they did not increase accuracy or convergence speed, the accuracy was much more stable in convergence and led to much more consistent results. Dropout layers had the opposite effect, which may be due to the relatively small amount of data used for this experiment that would not allow for individual weights to account for those that are dropped out.



Similar analysis was conducted for each separate model, and then the three models were combined into a multi-modal architecture based on those optimal hyperparameter choices. The resulting accuracy of each model is summarized below, as well as how long it took each model to converge:

Model	Accuracy	Time
MFCC CNN	68.80%	6.82 seconds
MFCC CNN & RNN	68.53%	47.71 seconds
DWT CNN	66.13%	1.14 seconds
STFT	65.87%	159.55 seconds
Multimodal	61.73%	170.12 seconds

As can be seen, all of the models did relatively the same in regards to performance. Because of this, there was not convincing evidence to show that a multimodal architecture was worth the extra time and computational expense; in fact the performance actually showed a noticeable drop in accuracy when all three features were combined. Rather, using one of the other models with significantly shorter training times like the MFCC model using just a CNN or the DWT model showed promising accuracy in a fraction of the amount of time. Therefore, for the use of personalized music recommendation on streaming platforms those two models would be of much more use to users.

The major drawback of the experiment described above was the scarcity of data used. Additionally, the use of mainstream songs possibly biased the data with confounding variables as many respondents commented that they liked a song at one point, but it had been ”overplayed” to the point they no longer liked it. The limitation of the sampling pool to CU college students also likely affected the results, though a manual review of the results still showed promising diversity of responses. Because commercial music platforms have access to an incredibly large amount of data from users of all backgrounds, as well as access to more than a single song to summarize a user’s taste, a lot more analysis can be done with systems like this and reach more conclusive results. Additionally, a lot of the hyperparameters explored would most likely have a much more meaningful impact on results if being trained with significantly more data. Overall though, the accuracy of the models demonstrated the potential use of models trained on digital audio signals if more user and song data was available for analysis.

5 Conclusions

The primary contribution of this project is the development and evaluation of a deep learning-based music classification and recommendation system that leverages digital signal processing techniques - such as MFCCs, STFT, and DWT - to extract meaningful audio features directly from raw music signals. By feeding these features into tailored CNN and CNN-RNN architectures, the model effectively captures both spatial and temporal structures in music. Experimental results show that single-modal models, especially those based on MFCCs and DWTs, achieve competitive accuracy with significantly less computational cost than a combined multi-modal architecture, indicating that simpler models may be more efficient for real-world music recommendation applications when data is limited.

While this approach advances personalized music recommendation by focusing on acoustic content rather than metadata or user behavior, it introduces potential ethical concerns around fairness, transparency, and social impact. First, reliance on mainstream music in the dataset may reinforce popularity biases and fail to represent diverse musical cultures or independent artists, potentially marginalizing underrepresented genres. Second, the use of user preference data, even anonymized, raises privacy concerns if scaled to real-world platforms without clear consent and data governance. Lastly, models trained on limited or skewed datasets may produce recommendations that reflect existing social or cultural biases, making it essential to ensure accountability and fairness in future iterations through broader, more inclusive datasets and transparent evaluation metrics. These concerns could be addressed by training models more exclusively on individual user preferences, which would also lead to a more personalized experience for each user.

References

1. Müller, M., Ellis, D.P.W., Klapuri, A., Richard, G.: Signal processing for music analysis. IEEE Journal of Selected Topics in Signal Processing **5**(6) (2011) 1088–1110
2. Mycka, J., Mańdziuk, J.: Artificial intelligence in music: recent trends and challenges. Neural Computing and Applications **37** (2024) 801–839
3. Zhang, Y., Li, T.: Music genre classification with parallel convolutional neural networks and capuchin search algorithm. Scientific Reports **15** (2025) 9580
4. Elbir, A., Aydin, N.: Music genre classification and music recommendation by using deep learning. Electronics Letters **56** (2020) 627–629
5. Schedl, M.: Deep learning in music recommendation systems. Frontiers in Applied Mathematics and Statistics **5** (2019)
6. Zhang, Y.: Music recommendation system and recommendation model based on convolutional neural network. Mobile Information Systems **2022** (2022) 1–14
7. Damak, K., Nasraoui, O., Sanders, W.S.: Sequence-based explainable hybrid song recommendation. Frontiers in Big Data **4** (2021)