AI-Based Genre Classification and Recommendation System for Nepali Music

Hari Prasad Baral
Department of Electronics and Computer Engineering, Pashchimanchal Campus, Institute of Engineering (IOE)
Pokhara, Nepal
haripbaral@wrc.edu.np

(Manuscript Received 5th May, 2025; Revised 20th May, 2025; Accepted 25th May, 2025)

Abstract

The music industry has experienced significant evolution with the emergence of digital technologies, especially through streaming platforms that have transformed how music is created, distributed, and consumed. As internet usage expands, music remains an essential aspect of daily life. Over the past three decades, traditional methods of discovering and listening to music have shifted drastically. Streaming services now provide instant access to vast music libraries, making them the preferred medium for listeners globally. This research project, Nepali Music Type Analysis and Recommendation, seeks to develop a Nepali-focused music streaming platform that enhances user engagement through intelligent music classification and personalized song recommendations. By leveraging machine learning to analyze user behavior and song attributes such as genre and style, the system aims to deliver customized music suggestions. In doing so, it not only improves music discoverability but also supports and promotes local Nepali artists through an intuitive and user-friendly platform.

Keywords—*Genre Classification, Music Recommendation System, Nepali Music, Artificial Intelligence, Streaming Platforms.*

1. INTRODUCTION

With the rise of internet accessibility, the way people interact with music has undergone a dramatic shift. Physical ownership of music has been replaced by digital consumption, with users increasingly turning to online streaming platforms for instant and convenient access to extensive music collections. Over the last few decades, this shift has influenced not only listening habits but also how new music is discovered.

Streaming platforms have become the dominant medium for music consumption, offering rich libraries and on-demand playback. However, one of the ongoing challenges is organizing and classifying large volumes of music efficiently. Manual genre classification is labor-intensive and error-prone, especially as the diversity of music grows. This has led to the adoption of machine learning techniques to automate genre classification based on audio features, metadata, and user interactions (Das et. al., 2014)..

Genre classification plays a critical role in music recommendation systems. Accurate genre labeling allows platforms to build smarter playlists, improve song suggestions, and personalize user experiences. Techniques such as Mel Frequency Cepstral Coefficients (MFCCs) have been widely used in music and speaker recognition due to their ability to capture timbral texture (Lindasalwa al., 2010). Further improvements have been proposed using harmonic-percussion separation and autoregressive modeling (Yuan et. al., 2015)., as well as hybrid approaches involving dynamic time warping (Abdalla et. al., 2010). and deep learning architectures like LSTM and autoencoders (Ghosal et. al, 2012; Logan, 2000).

1.1 Problem Statement

Despite the popularity of global music streaming platforms like YouTube and Spotify, Nepali users face significant limitations when accessing and exploring local music content. These platforms often lack dedicated systems for accurate classification of Nepali music by genre. As a result, users searching for specific types of Nepali songs must rely on user-generated playlists or irrelevant recommendations.

For example, while Spotify is available in Nepal, it does not provide genre-specific filters tailored

to Nepali music. YouTube, although accessible and widely used, restricts background playback for non-premium users and does not feature a structured recommendation system for local content. Additionally, services like YouTube Music are not officially available in Nepal, creating a gap in the domestic music streaming market.

Moreover, most current recommendation systems depend on basic metadata such as artist names or listening history, which may not effectively capture the nuanced preferences of Nepali users. These systems often result in mismatched or generic recommendations (Lindasalwa al., 2010). Manual classification of Nepali music, given its diversity, is also impractical and time-consuming (Das et. al., 2014). This highlights the need for an automated genre classification and recommendation system that caters specifically to Nepali music, enabling accurate song categorization and personalized listening experiences.

1.2 Objectives

The primary goal of the Nepal Tunes platform is to create a web-based application for Nepali music streaming that integrates genre-based classification and personalized recommendation features. The core objectives are:

- Develop a web-based Nepali music streaming platform with automatic genre classification using machine learning.
- Provide personalized song recommendations based on the genre of the currently playing track.

2. Literature Review

The music industry has experienced a significant transformation over the last two decades, largely driven by advancements in digital technologies and machine learning. As digital streaming platforms replace traditional methods of music consumption, the need for intelligent music classification and personalized recommendation systems has become increasingly important. This evolution has also impacted how users interact with music, emphasizing the need for accurate and scalable solutions, especially in culturally rich but underrepresented domains like Nepali music.

2.1 Music Genre Classification

Classifying music by genre is a central aspect of Music Information Retrieval (MIR). Traditional methods have mostly relied on metadata such as artist names, albums, and manually tagged genres. However, such manual approaches are limited in scalability and accuracy, especially when dealing with culturally diverse and expansive datasets like Nepali music (Das et. al., 2014). To overcome these issues, researchers have shifted toward audio-based techniques that analyze the sound signal itself.

Mel Frequency Cepstral Coefficients (MFCCs) are among the most widely used features for audio classification due to their effectiveness in capturing the timbral qualities of music (Lindasalwa al., 2010). Initially used for speech recognition, MFCCs have proven valuable in music genre classification as well (Rump et. al, 2010). Further improvements have been made by integrating MFCCs with other techniques (Yuan et. al., 2015). enhanced genre classification through harmonic-percussion separation, while (Abdalla et. al., 2010). used Dynamic Time Warping (DTW) to improve accuracy under noisy conditions. Wavelet-based approaches (Chen et. al., 2023). and deep learning models like BiLSTM with autoencoders (Ghosal et. al, 2012). offer promising advancements in genre classification, particularly for complex or non-Western music.

2.2 Recommendation Systems

Music recommendation systems are designed to suggest songs based on user behavior, preferences, and contextual information. Traditional approaches like collaborative and content-based filtering have been widely used, but they often perform poorly in regions like Nepal due to limited local data and differing musical tastes from global trends. Popular platforms such as YouTube and Spotify primarily rely on listening patterns, artist similarities, and user-created playlists, which may result in generic or culturally irrelevant recommendations (Lindasalwa al., 2010)..

To enhance relevance for Nepali listeners, a recommendation system must first accurately classify genres (Das et. al., 2014). introduced a hierarchical model using MFCC variants and amplitude patterns

to better distinguish similar genres—an essential feature for Nepali music, where genre boundaries can be subtle. Combining genre classification with user interaction data and deep learning techniques, such as those proposed by (Ghosal et. al, 2012). and (Abdalla et. al., 2010)., can lead to more personalized and context-aware recommendations that adapt to unique regional preferences.

2.3 Gaps and Need for Localized Solutions

Although significant advancements have been made in music classification and recommendation technologies, most existing systems cater to Western or mainstream global music. Nepali music remains underrepresented, with platforms like YouTube and Spotify lacking proper genre filters or structured support for local content. Additionally, limitations such as restricted background playback and the absence of services like YouTube Music in Nepal reduce usability. Given the vast diversity of Nepali music, manual tagging is inefficient. Therefore, there is a strong need for machine learning-based solutions to automate classification and deliver culturally relevant recommendations.

3. METHODOLOGY

3.1 Research Framework

The research framework for this study is divided into three core components: dataset preparation, music genre classification, and music recommendation. Each phase contributes to building a robust and intelligent music streaming platform focused on Nepali music.

3.2 Dataset Preparation

To carry out genre classification, a dataset consisting of 800 Nepali songs was compiled, evenly distributed across eight genres: adhunik, lok dohori, teej, pop, classical, rock, hip-hop, and gajal, with 100 songs per genre. This dataset was created by merging an existing collection of songs from four genres with manually sourced music from additional genres, ensuring diversity and balance.

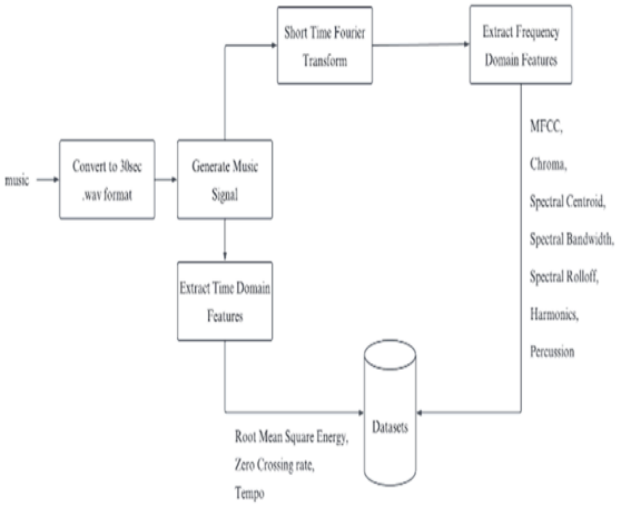

Fig. 1: Architecture of Data: Feature Extraction

To standardize the audio data, each song was trimmed or converted into a 30-second WAV file. This uniformity allowed for consistent feature extraction across all tracks. The audio features extracted include:

- Tempo
- Harmonics
- Mel-frequency cepstral coefficients (MFCC)
- Perceptual features
- Zero-Crossing Rate
- Spectral Roll-Off
- Spectral Bandwidth and Centroid
- Chroma features

- Root Mean Square Energy (RMSE)

These features serve as the foundation for both classification and recommendation.

3.3 Genre Classification

To classify music genres, multiple supervised machine learning algorithms were utilized. These include Decision Tree, Random Forest, Support Vector Machine (SVM), Naive Bayes, and others. The extracted features form a feature vector, which serves as the input for these models. The model outputs the predicted genre label, allowing the system to categorize songs efficiently.

3.4 Music Recommendation

To suggest similar songs to users, cosine similarity was employed as the similarity measurement technique. This algorithm compares the feature vector of a currently playing song with those in the dataset to identify tracks with the highest similarity. The output is a ranked list of songs that share musical characteristics with the input song.

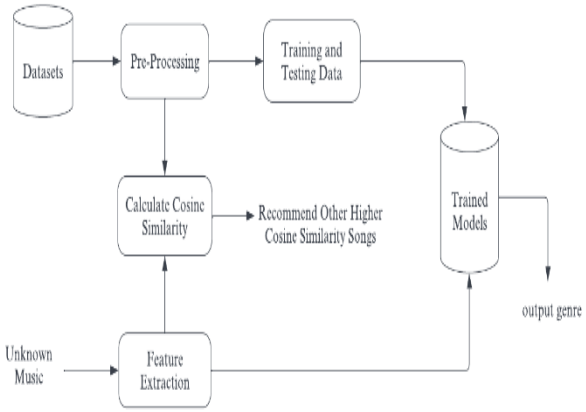

Fig. 2: Architecture of Genre Classification and Music Recommendation

4. Methods and Tools

The dataset used in this study was curated from both pre-existing collections and manually sourced Nepali songs. Initially, four genres—rock, lok dohori, nephop, and gajal—were collected from available repositories. Additional songs were then added manually to ensure genre diversity and reach the desired dataset size.

To ensure uniformity, all audio files were converted to 30-second WAV format. Feature extraction was carried out using the Librosa library in Python, which provided a comprehensive analysis of each audio sample's characteristics.

For classification tasks, the study employed a variety of supervised machine learning models including:

i. K-Nearest Neighbors (KNN)
ii. Support Vector Machine (SVM)
iii. Random Forest
iv. Gaussian Naive Bayes
v. XGBoost
vi. Decision Trees
vii. Logistic Regression

These models were implemented using Python's scikit-learn library. Cosine similarity was used to compute the resemblance between songs, allowing for content-based recommendations based on the extracted features.

4.1 Data Analysis Tools and Techniques

The study employed a variety of tools and techniques for data analysis, including feature extraction and classification. The Librosa package was utilized to extract key audio features such as Mel-frequency

cepstral coefficients (MFCCs), spectral contrast, and chroma features, which provide a representation of the audio signal's spectral and temporal characteristics. The scikit-learn library was then used to implement and evaluate different machine learning algorithms, including KNN, SVM, Random Forest, and others, for classification. The performance of each algorithm was assessed using confusion matrices and accuracy scores, providing insight into the classification accuracy across the genres. For a more thorough evaluation, additional metrics such as precision, recall, and F1-score could be considered, especially for multi-class classification scenarios where class imbalance may exist (McFee et. al., 2010)..

In terms of preprocessing, the initial datasets were scaled using the MinMaxScaler function from scikit-learn. This was done to normalize the feature values to a consistent range between 0 and 1, ensuring that no individual feature disproportionately influenced the models due to differences in scale (Pedregosa et. al., 2010).

4.2 Spectrogram Visualization

In this study, we generated spectrograms using the Short-Time Fourier Transform (STFT) to analyze how frequency components of audio signals change over time. The STFT of a discrete signal x[n] is computed as:

$$\mathrm{STFT}\{x[n]\}(m,\omega) = \sum_{n=0}^{N-1} x[n] \cdot w[n-m] \cdot e^{-j\omega n}$$

where:

x[n] is the input signal, w[n] is the window function, ω is the frequency axis, m is the time shift, and

N is the window length.

To extract features such as Mel-frequency cepstral coefficients (MFCCs), we mapped the frequency axis to the Mel scale using the formula:

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

where f is the frequency in Hz.

For the STFT, we used a hop length of 512 samples and a window size of 2048 samples.

Spectrograms generated from STFT are typically represented on a linear frequency scale, which may not align with human auditory perception. To better match human hearing, we converted the linear spectrogram to a logarithmic scale using the formula:

$$\mathrm{dB} = 20 \times \log_{10}(\mathrm{magnitude})$$

where the magnitude is the linear magnitude of the complex STFT value.

4.3 Features Extraction

4.3.1 Time Domain Features

In this study, we extracted several time-domain features directly from the audio signals to analyze their characteristics:

- **Root Mean Square (RMS) Energy**: This feature measures the average power (loudness) of the audio signal. It is calculated using the formula:

$$\text{RMS Energy} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x(n)^2}$$

where x(n) represents the amplitude of the signal at sample nnn, and NNN is the total number of samples.

- **Zero Crossing Rate (ZCR)**: This feature indicates the rate at which the audio signal's amplitude crosses the zero axis, changing from positive to negative or vice versa. It is computed as:

$$\mathrm{ZCR} = \frac{1}{N-1}\sum_{n=1}^{N-1} 1_{\{x(n)\cdot x(n+1)<0\}}$$

where 1 is an indicator function that equals 1 if its argument is true and 0 otherwise.

- **Tempo**: Tempo refers to the speed or pace of a given piece and is typically measured in beats per minute (BPM). It can be calculated as:

$$\text{BPM} = \frac{\text{Number of Beats}}{\text{Duration in Minutes}}$$

This metric provides insight into the rhythmic characteristics of the audio signal.

4.3.2 Frequency Domain Features

In this study, we extracted several key features from the audio signals to analyze their characteristics:

- **Mel-Frequency Cepstral Coefficients (MFCCs)**: We utilized 13 coefficients to represent the short-term power spectrum of the audio. This choice balances capturing essential spectral properties while mitigating the risk of overfitting, especially given the limited dataset.
- **Spectral Centroid**: This feature indicates the "center of mass" of the spectrum, providing a measure of brightness in the sound. It is calculated as:

$$\text{Centroid} = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)}$$

- Spectral Bandwidth: This measures the width of the spectrum, indicating the range of

$$\text{Spectral Bandwidth} = \sqrt{\frac{\sum_{k=1}^{N}(f(k)-C)^2 \cdot |X(k)|}{\sum_{k=1}^{N}|X(k)|}}$$

frequencies present. It is computed as:
where C is the spectral centroid

- Chroma Features: These capture the intensity of the 12 different pitch classes (semitones of the musical octave) present in the audio, providing insight into harmonic and melodic content.

- Spectral Roll-off: This is the frequency below which a specified percentage (typically 85%) of the total spectral energy is contained. It is calculated by finding the frequency fr such that:

$$\sum_{k=1}^{f_r} |X(k)| = 0.85 \times \sum_{k=1}^{N} |X(k)|$$

- Harmonics and Percussive Components: Harmonics are frequencies that are integer multiples of a fundamental frequency, contributing to the perceived pitch. Percussive components relate to the rhythmic elements of the sound.
Harmonic Frequencies=n×f1
where n is the harmonic number, and f1 is the fundamental frequency.

5 Analysis Results

The study found that the genre classification algorithms generally performed well, with their accuracy levels presented in a table. Among the methods analyzed, SVM demonstrated the highest accuracy. Additionally, the study observed that the spectrogram generated from the data appeared black due to the initial audio waves being in a linear scale.

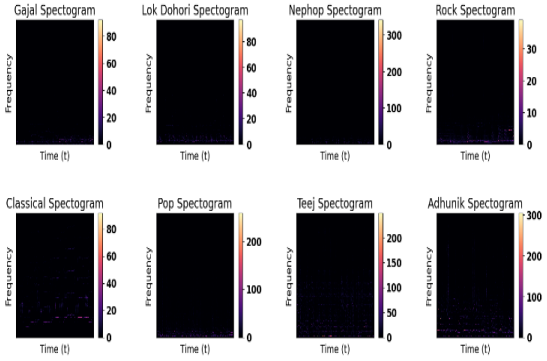

Fig. 3: Spectrograms of eight music genres

The image presents spectrograms of eight music genres, depicting frequency variations over time. The X-axis represents time, the Y-axis represents frequency, and color intensity indicates amplitude. Many spectrograms initially appeared dark due to the linear scale, which made low-energy frequencies less visible. To improve visualization, the study converted the spectrograms to a logarithmic scale (decibels, dB), aligning with human auditory perception. This adjustment enhanced the visibility of subtle frequency details, providing a clearer and more accurate representation. Further refinements, such as normalization or alternative color maps, could improve analysis.

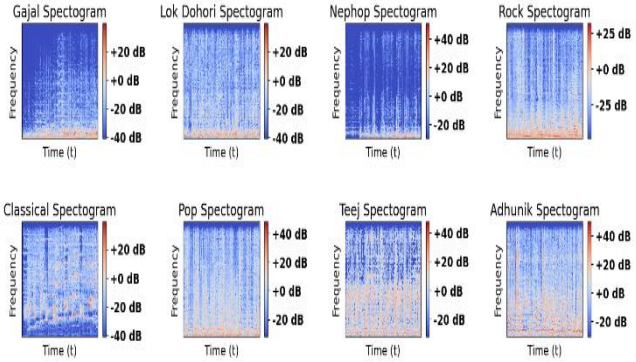

Fig. 4: Spectrograms of eight music genres [in dB]

For your music genre classification project, you used a 7:3 train-test split, with SVM achieving the highest accuracy at 93.33%, followed by Random Forest at 90.83%. Both models were evaluated using a confusion matrix, which provided insights into their performance. To enhance the results, you can try hyperparameter tuning for both SVM (e.g., adjusting kernel types and regularization) and Random Forest (e.g., optimizing n estimators and max_depth). Feature engineering can be improved by extracting more meaningful audio features like MFCCs and Chroma features, which could lead to better model performance.

Additionally, consider model ensembling by combining the predictions of SVM and Random Forest using a voting classifier to improve accuracy. For more reliable performance evaluation, k-fold cross-validation would help prevent overfitting and provide a more stable estimate of the model's performance. To enhance the dataset's diversity, use data augmentation techniques such as pitch shifting or adding background noise to increase the model's robustness. Finally, exploring deep learning models like CNNs or RNNs could provide even better results for audio classification.

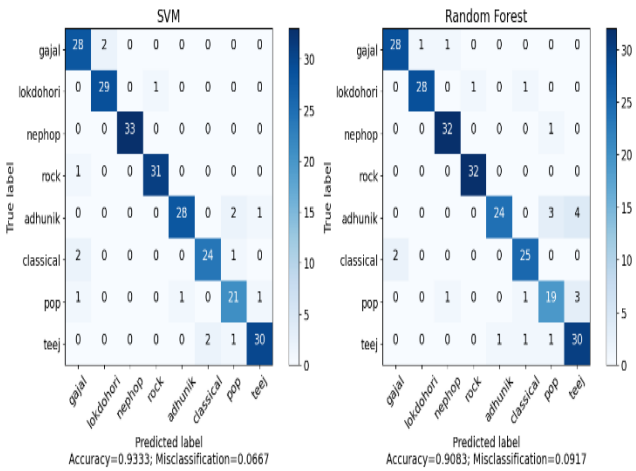

Fig. 5: SVM and Random Forest Prediction

The accuracy of each model is shown in the tables below.

Table 1: Accuracy of Various Models

SN	Algorithm	Accuracy
1	Gaussian Naive Bayes	86.67%
2	KNN	86.67%
3	Decision trees	75.00%
4	Random Forest	90.83%
5	Support Vector Machine	93.33%
6	Logistic Regression	87.50%
7	Multi-layer Perceptron	82.50%
8	Cross Gradient Booster	88.33%
9	Cross Gradient Booster (Random Forest)	86.67%

Music genre classification tests, including MFCC (Mel-frequency cepstral coefficients) in the dataset resulted in an accuracy of 93.33% for the SVM model. However, when MFCC was removed, the accuracy dropped to 81.67%, showing how important MFCC is for the model's performance. MFCC captures essential audio features such as pitch, timbre, and spectral content, which help the model distinguish between different genres effectively. Without these features, the model struggles to classify genres accurately.

Given the significant improvement with MFCC, it's advisable to continue using these features in your dataset. Additionally, you might explore combining MFCC with other audio features like Chroma features and Spectral contrast to enhance the model's understanding of the audio data. You could also experiment with deep learning models, which can automatically learn complex patterns from the audio for even better results.

6 Conclusion

This study focused on classifying music genres and recommending songs based on extracted audio features using machine learning methods. A dataset of 800 songs spanning 8 genres was analyzed, and multiple audio features—including tempo, harmonics, MFCC, perceptual features, zero crossing rate, spectral roll-off, spectral bandwidth, spectral centroid, chroma, and root mean square energy (RMSE)—were extracted. These features were then used as inputs for various supervised machine learning models such as Decision Trees, Random Forest, Support Vector Machine (SVM), and Naïve Bayes. SVM outperformed other models with a classification accuracy of 93.33%, followed by Random Forest at 90.83%. For song recommendations, cosine similarity was utilized to suggest tracks with similar audio characteristics

7 Further Works

This study can be expanded in several ways to improve music genre classification and recommendation. One approach is to increase the dataset's diversity by including a wider range of music genres. Additionally, integrating unsupervised machine learning algorithms could provide new insights into genre classification and recommendation. More advanced recommendation techniques beyond cosine similarity can also be explored to enhance accuracy. Furthermore, combining both audio and non-audio features, such as metadata and user preferences, could improve classification and recommendation, particularly for Nepali songs. Overall, this study has shown promising results, and further research could significantly enhance music discovery and user experience.

References

Abdalla, M. I., & Ali, H. S. (2010). Wavelet-based Mel-frequency cepstral coefficients for speaker identification using hidden Markov models. arXiv. arxiv.org

Chen, X., Pun, C. M., & Chen, C. L. P. (2010). Robust MFCC feature detection and dual-tree complex wavelet transform for digital audio watermarking. Information Sciences, 298, 159–179. https://doi.org/10.1016/j.ins.2014.11.040 research.mpu.edu.mo

Chen, Y.-L., Wang, N.-C., Ciou, J.-F., & Lin, R.-Q. (2023). Combined bidirectional long short-term memory with Mel-frequency cepstral coefficients using autoencoder for speaker recognition. Applied Sciences, 13(12),

7008. https://doi.org/10.3390/app13127008 ijcaonline.org+14mdpi.com+14grafiati.com+14

Das, A., Jena, M. R., & Barik, K. K. (2014). Mel-frequency cepstral coefficient (MFCC) – a novel method for speaker recognition. Digital Technologies, 1(1), 1–3. https://doi.org/10.12691/dt-1-1-1 joiv.org+5researchgate.net+5pmc.ncbi.nlm.nih.gov+5pubs.sciepub.com+1grafiati.com+1

Ganchev, T., Fakotakis, N., & Kokkinakis, G. (2005). Comparative evaluation of various MFCC implementations on the speaker verification task. In Proceedings of SPECOM 2005 (Vol. 1, pp. 191–194).

Ghosal, A., Chakraborty, R., Dhara, B. C., & Saha, S. K. (2012). Music classification based on MFCC variants and amplitude variation pattern: A hierarchical approach. International Journal of Signal Processing, Image Processing and Pattern Recognition, 5(1), 131–? researchgate.net+15grafiati.com+15zh.wikipedia.org+15researchgate.net

Lindasalwa Muda, Mumtaj Begam, & Elamvazuthi, I. (2010). Voice recognition algorithms using Mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW). arXiv. pmc.ncbi.nlm.nih.gov+7arxiv.org+7arxiv.org+7

Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. In Proceedings of the International Symposium on Music Information Retrieval (ISMIR). Retrieved from http://ismir2000.ismir.net/papers/logan.pdf

McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in Python. In S. van der Walt & J. Millman (Eds.), Proceedings of the 14th Python in Science Conference (pp. 18–25).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Rump, H., Miyabe, S., Tsunoo, E., Ono, N., & Sagayama, S. (2010). Autoregressive MFCC models for genre classification improved by harmonic-percussion separation. In Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR). ismir2010.ismir.net

Yuan, X. C., Pun, C. M., & Philip Chen, C. L. (2015). Robust Mel-frequency cepstral coefficients feature detection and dual-tree complex wavelet transform for digital audio watermarking. Information Sciences, 298, 159–179. https://doi.org/10.1016/j.ins.2014.11.040 research.mpu.edu.mo