The Impact of Audio Features on Music Genre Classification and Recommendations

Evaluating Audio Feature Importances and Machine Learning Models to Enhance Music Genre Classification and Recommendations

supervised by David Gauthier

Universiteit Utrecht

David Sijbesma, 6278035, d.t.sijbesma@student.uu.nl
Utrecht University

Abstract

Established music distributors like SoundCloud and Spotify use song recommendations to improve the user experience. These recommendations are typically based on user listening history, metadata, and collaborative filtering techniques. However, audio level features might also be a way to enhance the recommendation system. Additionally, these features provide the ability to further understand and classify genres, as genres are often misclassified by distributors like Spotify and SoundCloud. This misclassification can be attributed to human perceptions that do not align with the broader user consensus. By leveraging audio features, the consistency of these human-classified genres can be tested and potentially improve recommendation accuracy. Therefore, this research investigates the impact of various audio features on the accuracy of music genre classification and recommendation systems. A self-generated dataset and advanced machine learning models are used. The dataset, created using audio tracks from various genres, features Mel-Frequency Cepstral Coefficients (MFCCs), spectral centroid, chroma features, and zero-crossing rate. Five machine learning models—K-Nearest Neighbors (KNN), Random Forest (RF), Gradient Boosting (GB), Support Vector Machines (SVM), and eXtreme Gradient Boosting (XGB)—were utilized. Additionally, an ensemble model was also created using a stacking method to combine these base models with a meta-classifier. The most important findings indicate that MFCCs are most important in capturing timbral characteristics, while a combination of timbral, rhythmic, and pitch content features further enhances classification performance. Another important improvement is found by using an ensemble learning methodology, especially stacking models that combine several machine learning techniques to achieve better accuracy. The study also highlights the impact of dataset consistency and size on classification accuracy and the need for refined methods to differentiate genres with overlapping audio features. This research is an indication of what music classification and recommendation systems could contribute to the improvement of the user experience. Therefore, this research forms a solid basis for further studies to focus on improved genre classification, for instance by using deep learning models or a more extensive range of audio features. Ultimately, to further understand whether the recommendations using audio features contribute to the user experience, it is suggested to test this system with human feedback.

Keywords: Music Genre Classification, Audio Features, Music Recommendation, Machine Learning, Mel-Frequency Cepstral Coefficients (MFCCs), Ensemble Learning, Content-Based Recommendation System, Cosine Similarity

1. Introduction

Music has been an integral aspect of human entertainment and pleasure since ancient times, with significant social and psychological effects (Tzanetakis & Cook, 2002). With the internet as a major source of music distribution and sharing, various sites are dedicated to distributing and commercializing music. Platforms such as, SoundCloud and Spotify, along with widespread internet integration, make it easy for individuals to find music tailored to their preferences. These platforms typically offer recommendations based on listening habits, metadata, billboard rankings, trends, and the listening habits of others with similar taste (Newsroom, 2023). On this note, research has been done on contextual approaches to music, based on information related to music that is not based upon the audio components itself (Rebelo et al., 2012), such as music scores, keywords, metadata, playlists, user reviews, and others (Knees & Schedl, 2013). While these recommendations from these channels already provide high user- experience, they do not include the content of a song (i.e. the way a song is constructed), to improve these recommendations. To tighten this gap, researchers used methods based on the analysis of audio content (Pampalk et al., 2002; McKay & Fujinaga, 2009; Song et al., 2012). They found that the most basic form of music can be described by their pitch, loudness, and timbre, which includes amplitude envelope, harmonicity, and spectral envelope (which are considered in this paper as audio-level features), and can be used for music classification and retrieval (Wold et al., 1996).

The popularity of music has motivated the development of computational music analysis methods over the past decades (Burred & Lerch, 2004). A crucial element in this endeavor is Music Information Retrieval (MIR), which is the extraction of meaningful information from music files for purposes such as classification, recommendation, and analysis (McKay & Fujinaga, 2009; Song et al., 2012). Despite significant progress in MIR over the past decade, music recommendation systems are still in an early developmental stage (Afchar, 2023). There is limited

knowledge about how audio-level features affect the accuracy of genre classification and music recommendations. Currently, music genre classification is often performed manually (Tzanetakis & Cook, 2002).

Genres are defined by both cultural and linguistic factors, helping people understand and categorize different types of music, based on shared characteristics. This shared understanding is influenced by cultural and ideological conventions, making genres both flexible and dynamic (Heikkinen, 2012). They function as cognitive-semantic schemas that encompass ideational, interpersonal, and textual levels of meaning, allowing individuals to recognize and utilize them based on their linguistic competence. It has been observed that audio signals within the same genre mainly share specific characteristics, such as similar instruments, rhythmic patterns, and pitch distributions (West & Cox, 2004). This suggests that it should be possible to perform automatic music genre classification. However, the overlapping nature of genres, which evolve and influence each other over time, makes it harder to distinguish between them.

This article will therefore attempt to answer the following research question:

> *How do the relationships between individual audio features within music genres affect the accuracy of genre classification and recommendation systems?*

To answer this question, multiple analyses will be applied to audio features within the dataset to identify underlying patterns and improve classification accuracy and recommendation quality. By highlighting the most impactful audio features within each genre, this research aims to clarify intrinsic relationships between audio features within different genres and develop a meta-model that enhances both classification accuracy and the quality of genre-based recommendations. Furthermore, this analysis will reveal the influence of individual audio features and provide a framework for optimizing these features to improve personalized music recommendation systems.

2. Theoretical Background

2.1. Audio Features

Audio features are attributes or properties extracted from audio signals that provide meaningful information for various tasks such as MIR, speech recognition, and audio segmentation. These features help in creating a compact and expressive description of the audio or sound that is machine-processable (Mitrović et al., 2010). Audio features can be categorized into three distinctive levels: low-level, mid-level, and high-level audio features. Low-level audio features include short-term features such as spectral centroid, spectral rolloff, spectral flux, zero crossings, and Mel-Frequency Cepstral Coefficients (MFCC). These features are computed for short-time frames (e.g., 20-30 ms) and capture the immediate characteristics of the audio signal (Müller, 2015).

Mid-level audio features offer a more abstract representation, bridging the gap between low-level signal characteristics and high-level perceptual attributes. Examples include articulation and melodiousness, which capture qualities like the clarity of musical notes and the pleasantness of a sequence of sounds (Aljanaki & Soleymani, 2018).

High-level audio features, on the other hand, include long-term features such as rhythm, beat, and tempo, which capture the structural information of the audio over larger time scales. These features often align with human perception and contextual understanding. High-level features are often linked to metadata such as genre and social tags, providing a broader context that aligns with user preferences and the overall listening experience (Kim et al., 2020; Melchiorre & Schedl, 2020; Knees & Schedl, 2016).

Former research shows that several studies have examined different audio feature sets for classifying music genres. Tzanetakis & Cook (2002) found that using three main feature sets: timbral texture, rhythmic content, and pitch content gave an accuracy of 61% for ten musical genres. Another research by Chathuranga & Jayaratne (2013) focused on feature extraction from frequency, temporal, cepstral, and modulation frequency domains. They achieved an accuracy of up to 81% on the ISMIR2004 genre dataset using support vector machines (SVM) with a polynomial kernel. Subsequently, Burred & Lerch (2004) highlighted the importance of hierarchical classification, noting that feature selection should consider genre dependency. They implemented a hierarchical system to classify audio signals into 17 classes, achieving high classification accuracy.

2.2. Impact and Potential of Individual Audio Features

In the context of music classification, knowing the impact of individual audio features is of high importance. Research has identified that specific features such as the sum of the beat histogram and the variance of the spectral centroid are highly effective for genre classification. Subsequently, Chathuranga & Jayaratne (2013) noted that feature selection significantly influences classification accuracy, recommending the use of wrapper and filtering methods to identify the best features. Among these, Mel-frequency cepstral coefficients (MFCCs) have been particularly notable for their effectiveness in capturing timbral characteristics of audio signals (Abeysinghe et al., 2021)(Chathuranga & Jayaratne, 2013). Designed to mimic the human ear's response, MFCCs provide a compact representation of the spectral envelope, crucial for distinguishing between different timbres. Similarly, features like spectral flux and zero crossings offer insights into the noisiness and dynamic changes within the audio signal, which are essential in differentiating genres reliant on these characteristics.

Therefore, it is assumed that audio features, as outlined above, are an influential factor in the classification of genres and genre-based recommendations. From this, the following two hypotheses can be derived

- **Hypothesis 1:** Using only MFCCs can effectively classify music genres by capturing essential timbral characteristics.
- **Hypothesis 2:** Including timbral, rhythmic, and pitch content features significantly improves the accuracy of automatic music genre classification compared to using only MFCCs.

2.3. Comparison with Human Classification

Human classification of musical genres involves listeners using both objective measures and subjective perceptions to categorize music. Lippens et al. (2004) conducted experiments comparing human and automatic musical genre classification, finding that although automatic systems showed promising results, they still lagged behind human performance. Their experiments revealed the inherent subjectivity in genre classification and the challenges in achieving high accuracy with automatic systems.

As previously discussed, individuals have their own perspectives on genres, and genres often overlap due to cultural influences Heikkinen (2012). This raises the question of whether using a dataset that is consistent with a single human classifier's perspective yields better accuracy compared to a more diverse dataset defined by multiple people's genre classifications. The GTZAN dataset used by Tzanetakis & Cook (2002) was constructed and classified by a single person, resulting in a consistency from one individual's perspective, and achieved an accuracy of 61%.

However, modern tracks found on music platforms like Spotify and SoundCloud, are classified by humans but lack the consistency of a single individual's viewpoint. This variability arises because not everyone shares the same opinion on what defines a genre, reflecting cultural differences and the overlapping nature of genres West & Cox (2004). Additionally, these platforms also allocate many sub-genres, which dilutes the applicability of solely one genre to a track, and creates inconsistencies.

Therefore, the following hypothesis is proposed:

- **Hypothesis 3:** Using a more consistent dataset (e.g., GTZAN Dataset) results in higher accuracy for music genre classification by machine learning models compared to a dataset classified by multiple people (e.g., a Spotify-extracted dataset).

Moreover, the size of the dataset could also be of influence. Modern music platforms are continually expanding, encompassing numerous sub-genres. It is reasonable to assume that a larger dataset could improve accuracy (given that the dataset remains consistent (Kavzoglu, 2009), as machine learning models would have more data to learn from and could better identify underlying patterns in music, including sub-genres. This leads to the next hypothesis:

- **Hypothesis 4:** Increasing the size of a dataset leads to better genre classification accuracy compared to a smaller dataset.

2.4. Machine Learning Techniques for Music Genre Classification

Deep learning methods, particularly convolutional neural networks (CNNs), have shown exceptional promise in extracting high-level features from audio data. Elbir & Aydin (2020) demonstrated a CNN-based system for genre classification and music recommendation, achieving high accuracy by processing mel-spectrogram images. Deep learning models, while highly effective, often function as "black-box" models (Martínez Ramírez et al., 2020). This means that their internal workings are not easily interpretable, making it challenging to understand why certain decisions are made, despite their high performance (Elbir & Aydin, 2020).

Opposed to deep learning methods, machine learning techniques have also proven to be highly effective for music genre classification. Among these, support vector machines (SVMs) are renowned for their robustness in handling high-dimensional spaces, making them suitable for complex audio data (Li et al., 2003). K-nearest neighbors (k-NN) offer simplicity and effectiveness, particularly for smaller datasets, by leveraging the proximity of data points in the feature space. Decision trees and their ensemble variants, such as random forests, enhance classification performance by reducing overfitting and improving generalization through the aggregation of multiple decision paths.

Additionally, research by Ahlem et al. (2022) & Li et al. (2003), found that stacking machine learning models into an assembled classifier can enhance the system's ability to accurately define or classify the genre and handle complex and diverse user preferences. A stacking model aggregates the predictions of several base classifiers (e.g., SVM, k-NN, random forest) and uses a meta-classifier to make the final prediction. This approach leverages the diverse strengths of each algorithm, potentially leading to superior overall performance compared to individual models.

Therefore, it is assumed that utilizing various machine learning techniques can have a positive effect on music genre classification:

- **Hypothesis 5:** Utilizing multiple machine learning techniques in a single stacking model will improve the accuracy of genre classifications.

2.5. Audio Features and Music Recommendations

As previously mentioned, music recommendations are often based on metadata rather than underlying audio features. This raises the critical question of whether music recommendations can be effectively made based solely on audio-level features, without relying on metadata such as predefined genres.

The traditional approach of using metadata for recommendations assumes that genres are clearly defined and mutually exclusive. However, genres frequently overlap and evolve, influenced by cultural and linguistic factors, making strict categorization challenging Heikkinen (2012). This overlap can result in recommendations that are limited to predefined genre boundaries, potentially overlooking tracks that share similar audio characteristics but belong to different genres. West & Cox (2004) observed that audio signals within the same genre often share specific characteristics, such as similar instruments, rhythmic patterns, and pitch distributions, suggesting that it might be feasible to base recommendations on these audio-level features.

However, another important aspect of this study is to research whether models can give accurate recommendation based solely on audio features. Given the overlap in audio features across different genres, recommendations might inherently include tracks from various genres, leading to a more diverse and potentially more satisfying listening experience. This approach contrasts with metadata-based recommendations, which may be restricted to tracks within the same genre.

To explore this, the following hypothesis is proposed:

- **Hypothesis 6:** Music recommendations based solely on audio-level features may provide more diverse recommendations that include tracks from different genres due to the overlap in audio characteristics.

2.6. Summary

The theoretical background explores the classification and recommendation of music genres using audio features and machine learning techniques. Audio features are categorized into low-level, mid-level, and high-level, capturing various characteristics of audio signals. Existing literature reveals gaps in understanding the impact of dataset consistency and size on classification accuracy, and the potential of using only audio features for music genre classification and recommendations without relying on metadata. This research aims to address these gaps by comparing the accuracy of genre classification using consistent datasets, like GTZAN, versus more diverse datasets, such as those extracted from Spotify. Additionally, the study will examine how dataset size affects classification performance and explore the use of stacking models that combine multiple machine learning techniques to enhance accuracy. Lastly, it will evaluate the feasibility of generating diverse and satisfying music recommendations based solely on audio features, aiming to improve the precision and user satisfaction of music classification and recommendation systems.

3. Data & Feature selection

In order to answer the research question, "How do the relationships between individual audio features within music genres affect the accuracy of genre classification and recommendation systems?", a self-generated dataset was used. The inspiration for the audio feature selection used in this paper comes from the GTZAN Genre Collection dataset. This dataset is used in multiple researches for genre classification and is recognized as the "MNIST of sounds" within the music information retrieval (MIR) community (Li et al., 2023). It contains 1,000 audio tracks evenly distributed across 10 musical genres, with each track being a 30-second sound clip. Based on this dataset, a new, extensive dataset (containing more than 5000 audio tracks) has been created. The new dataset maintains the same audio features and 30-second sound clips, but another composition of genres was selected.

Apart from the self-generated dataset, the GTZAN dataset created by Tzanetakis & Cook (2002) was also used for the comparison between consistency in human classification.

3.1. Extraction

These audio features and sound clips are extracted using the Spotify API, utilizing Python code and the librosa package. The librosa package is a widely used tool in MIR, offering great flexibility to expert users (Downie, 2003). Additionally, the Spotify API is employed to extract album cover paths for album visualizations in the recommender system interface that is created later on. Feature extraction involves computing a compact numerical representation that characterizes a segment of audio. Designing descriptive features for a specific application is the primary challenge in building pattern recognition systems.

3.2. Levels of Audio Features

Understanding the distinction between low-level, mid-level, and high-level audio features is crucial for genre classification. This section categorizes the various audio features used in this study into these levels, highlighting their importance in representing different aspects of audio signals.

3.2.1. Low-Level Audio Features

Low-level audio features capture the immediate characteristics of the audio signal, typically computed for short-time frames (e.g., 20-30 ms). These features include:

- **Rhythmic Content Features**
  - **Spectral Centroid:** Measures the center of mass of the spectrum and indicates the brightness of the sound (Tjoa, 2017).
  - **Spectral Bandwidth:** Reflects the width of the spectral energy distribution.
  - **Spectral Rolloff:** The frequency below which a certain percentage (usually 85%) of the total spectral energy is contained, indicating the spectral shape (Tjoa, 2017).
  - **MFCCs (Mel-Frequency Cepstral Coefficients):** Captures short-term spectral-based features by transforming the logarithm of the power spectrum into the mel scale, followed by a discrete cosine transform to decorrelate the coefficients (Abeysinghe et al., 2021). The coefficients used are MFCC 1 to MFCC 13 .

3.2.2. Mid-Level Audio Features

Mid-level audio features offer a more abstract representation, bridging the gap between low-level signal characteristics and high-level perceptual attributes. These features include:

- **Rhythmic Content Features:**
  - **Zero Crossing Rate:** The rate at which the signal changes sign, indicative of the noisiness of the signal.
  - **Spectral Contrast:** Measures the difference in amplitude between peaks and valleys in a sound spectrum, calculated across five different bands (Jiang et al., 2002).
  - **RMS Energy:** The root mean square of the signal's amplitude, reflecting the energy content of the audio.
- **Pitch Content Features:**
  - **Chroma Features:** Represent the twelve different pitch classes, (C, C#, D, D#, E ,F, F#, G, G#, A, A#, B) (Ellis, 2007) providing information about the harmonic content of the music. These are calculated from Chroma STFT 1 to Chroma STFT 12.
  - **RMS Harmonic:** Root mean square of the harmonic components of the audio signal.
  - **RMS Percussive:** Root mean square of the percussive components of the audio signal.

3.2.3. High-Level Audio Features

High-level audio features capture the structural information of the audio over larger time scales, often aligning with human perception and contextual understanding. These features include:

- **Dynamics and Texture Features:**
  - **Tempo (BPM):** Beats per minnut (BPM) is the overall speed or pace of a piece of music, calculated from the time-domain amplitude envelope.
  - **Rhythm:** Includes beat and tempo, which capture the structural information of the audio over larger time scales.

4. Methodology

This study utilizes the CRISP-DM (Cross-Industry Standard Process for Data Mining) model to systematically develop an automated music recommendation system (Chapman et al., 2000). As a commonly used model in data mining and machine learning projects, CRISP-DM provides a structured approach with six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. As visualized in Figure 1, this research employs the CRISP-DM methodology to concentrate on the early stages of research and data understanding. This is followed by thorough data cleaning, feature importance analysis, and the development of classification models, culminating in the implementation of a content-based recommendation system.

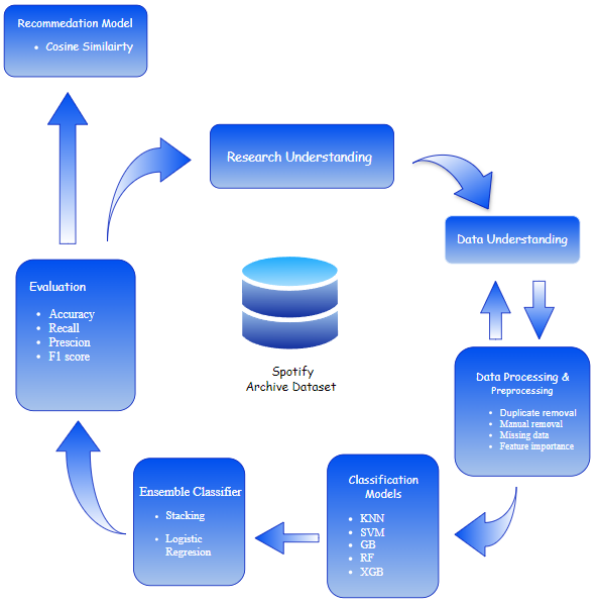

Figure 1: CRISP-DM Process Model

4.1. Research Understanding

This research explores the impact of audio features on genre classification and music recommendations. It aims to enhance personalized recommendations by analyzing how specific audio features influence accuracy. Utilizing machine learning, the study develops an automated recommendation system with an interactive interface for user input and track matching. The goal is to improve classification accuracy and recommendation quality through detailed feature importance analysis.

4.2. Data Understanding

In this research, a new large-scale dataset has been created, preserving the same audio features and the 30-second sound clip format of GTZAN. Using the Spotify API and the Python package librosa, 30-second MP3 files from specified genres are extracted. These genres are: blues, classical, country, disco, hip-hop, jazz, metal, techno, and reggae. The dataset consists of 5000+ songs. Each track in the dataset is stored in a database along with its extracted audio features, MP3 file, and album cover. The inclusion of album covers is to provide a more engaging user experience in the interface.

4.3. Data Cleaning, selection and preprocessing

A wide set of features was extracted using the librosa package, which includes the following: chroma features, spectral features, and Mel-frequency cepstral coefficients (MFCCs). These were then standardized by a StandardScaler, which ensures that all of the features contribute equally. Genre labels were encoded into numerical values using LabelEncoder to be fed into machine learning algorithms.

Class imbalance was handled through the balancing of the dataset by down-sampling each genre to the size of the smallest class, so that it would constitute an equal share to prevent bias. The dataset, which contains over 5000 tracks, each having 41 individual audio features, was split into an 80-20 ratio for training and testing to achieve the best performance (Muraina, 2022). Cross-validation was performed on the models to ensure robust performance and to avoid overfitting. For model training, the predictions from KNN, Ensembled Random Forest, Gradient Boosting, eXtreme Gradient Boosting, and SVM base models were used in making meta-features that would be utilized in training an assembled model with logistic regression for the purpose of stacking. As such, the ensemble approach utilizes the strengths of every base model in improving classification accuracy.

In the process of cleaning data, all mismatches that base models found were particularly analyzed. This step allowed for the analysis of the subsets of songs that every model was consistently misclassifying to identify possible errors in genre labeling on Spotify API. Every song in those subsets was manually reviewed to check for errors in genre. Those with incorrect genre labels were removed from the dataset to maintain integrity and to enhance the reliability of further model training.

4.4. Feature Importance & Classification Models

To establish and quantify the contribution of each feature in making predictions by a model a technique is used to find the feature importances. This helps to understand which features are most informative for the target variable and hence allows the derivation of underlying patterns within the data. The research will therefore establish the importance of different audio features on genre classification.

4.4.1. Neural Network for Feature Importance

To this end, a Multilayer Perceptron (MLP) classifier will be used, which is one kind of feedforward artificial neural network. MLPs are proven to capture complex, nonlinear relationships in the data. The MLP is particularly well-suited for high-dimensional feature spaces, such as those encountered in audio classification (Lin et al., 2007). The steps involved in this analysis are as follows:

1. **Data Preparation:** The numerical columns should be standardized to have a mean of zero and a standard deviation of one. This normalization will ensure all features are treated equal in the learning by the model (to prevent bias).
2. **Binary Target Creation:** A binary target variable is created wherein the genre of interest is labeled as 1 and all other genres as 0. Thus, this multi-class classification problem would get restructured into multiple binary classification problems, where it now has the chance to focus on only one genre at a time.
3. **Model Training and Evaluation:** Here, the dataset is divided into a training subset and a test subset, and balanced representation is done. The MLP classifier is trained using the training data. Permutation importance is used by shuffling the values of each feature and measuring the decrease in model accuracy for the calculation of the feature importance; a large drop in pertains high importance to a feature.

In the analysis, according to permutation importance values, top 10 permutations are selected for every genre. For comparison, their importance values are then normalized and plotted for visualization.

4.4.2. Classification models

This section provides a detailed overview of five machine learning classifiers used in this research study and an ensemble method implemented for improving the predictive performance. The classifiers include K-Nearest Neighbors (KNN), Random Forest (RF), Gradient Boosting (GB), XGB, and Support Vector Machines (SVM). Additionally, an ensemble model using a meta-classifier is discussed.

1. K-Nearest Neighbors (KNN). K-Nearest Neighbors (KNN) is a simple, but effective, instance-based learning algorithm used for classification. The fundamental idea is to predict the class of a data point by examining the classes of its nearest neighbors in the feature space. The number of neighbors, k, is a crucial hyperparameter, and in this study, k=5 is chosen. KNN is highly intuitive and works well with smaller datasets where the decision boundary is not very complex. It is a non-parametric method, meaning it makes no underlying assumptions about the data distribution (Guo et al., 2003).

2. Random Forest (RF). Random represents an ensemble learning method that synthesizes predictions from a predefined number of decision trees. It operates based on two core principles:

1. **Bootstrap Aggregation (Bagging)**: Each decision tree is trained on a random subset of the training samples, known as bootstrap aggregation or bagging (Breiman, 1996).
2. **Random Feature Subsets**: Each decision tree makes its predictions using only a random subset of the features (Amit & Geman, 1997).

The final predicted class in a Random Forest is determined by the majority vote of the individual trees. This methodology enhances model robustness and reduces overfitting by averaging the predictions of multiple decision trees.

3. Gradient Boosting (GB). Gradient Boosting is another ensemble learning method that integrates many weak learners, like decision trees, into a stronger predictive model. Unlike random forests, where individual trees are trained independently, boosting algorithms train trees sequentially using forward stagewise additive modeling (Hastie et al., 2009). The early iterations will produce relatively simple decision trees, but as the training progresses, the model pays extra attention to those instances on which the previous learners make errors. The final prediction is a weighted linear combination of outputs from individual learners.

4.Extreme Gradient Boosting (XGB). XGB is an advanced implementation of gradient boosting for both speed and performance. It brought regularization techniques to prevent overfitting, in addition to parallel tree boosting in XGBoost for efficient computation. In particular, it included significant features: built-in handling of missing values, fleshing out a variety of objective functions, etc., making it highly versatile in various classification task genres. XGB was selected because big datasets are extended capacities to handle them and return robust and accurate predictions in this study.

5. Support Vector Machines (SVM). Support Vector Machines are a set of supervised learning methods used for classification that transform the original input data into a higher-dimensional space using a kernel trick (Cortes & Vapnik, 1995). In this transform space, data will be linearly separated with the use of a hyperplane. The optimal hyperplane among such categories is one that maximizes the margin between classes of different nature. In this regard, in this research, an RBF kernel SVM was used, since it would be much more appropriate for modeling nonlinear relationships. Similar to logistic regression, SVM is actually implemented as a one-vs-rest classification task so that it can effectively handle the multi-class classification problem.

4.5. Ensemble Model: Meta-Classifier

Besides these individual classifiers, a meta-classifier was implemented that seek to incorporate the strengths of these base models. The meta-classifier uses the outputs from the base models—in this case, KNN, Random Forest, Gradient Boosting, XGB, and SVM—as input features in training a logistic regression model. This approach is what is referred to as stacking, which allows the meta-model to learn the best way to integrate these base models with their predictions to achieve better outcome. The steps involved in training and evaluating the meta-classifier are as follows:

1. **Train Base Models:** Each base model (KNN, RF, GB, XGB, SVM) is trained on the scaled training data.
2. **Generate Meta-Features:** Predictions (probabilities) from each base model are generated for both the training and test datasets.
3. **Train Meta-Model:** A logistic regression model is trained on the meta-features generated from the base models' predictions on the training data.
4. **Evaluate Meta-Model:** The trained meta-model is used to make final predictions on the test dataset, combining the strengths of all base models to achieve higher accuracy and robustness.

This ensemble approach could enhance predictive performance by leveraging the complementary strengths of different classifiers, leading to more accurate and reliable genre classification.

4.6. Content-Based Recommendation Model

Content-based recommendation systems suggest items similar to those the user has previously interacted with by leveraging the item's content. This method is particularly effective in scenarios with sparse user data, addressing the cold-start problem by using the inherent properties of the items themselves(Darshna, 2018). In this study, a content-based recommendation system is implemented using cosine similarity, This is a widely recognized metric that measures the angle between two feature vector, by doing so it provides a measure of closeness between them (Sheikh Fathollahi & Razzazi, 2021). It is computed as the dot product of the vectors divided by the product of their magnitudes, as shown in the equation below:

$$\cos(x, y) = \frac{x \cdot y}{\|x\|\|y\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} (x_i)^2} \sqrt{\sum_{i=1}^{n} (y_i)^2}}$$

The system evaluates the similarity between the feature vectors of an anchor track provided by the user and other tracks in the dataset. Genre prediction is performed by first extracting features from the uploaded audio file and scaling them using the pre-fitted StandardScaler. The extracted features are then used to obtain probability predictions from each of the base classifiers. These probability predictions are concatenated into one feature vector and fed into the meta-classifier to output the final genre prediction. This is returned to the user. Then, the computation of cosine similarity between the feature vectors of the uploaded audio and those of all tracks in a dataset will be done for identifying the most similar tracks. It then shows the user, together with their corresponding audio files and album covers, the top 10 tracks selected according to their largest acquired similarity scores.

5. Architecture Design

5.1. User Interface Layer

The User Interface Layer operates through Streamlit, where users upload MP3 files for analysis. This layer provides an intuitive interface for users to interact with the system, upload audio files, and receive genre predictions and recommendations.

5.2. Data Persistence Layer

This layer is responsible for sourcing MP3 files and audio features via the Spotify API and processing them with the librosa library. Cleaning the data involves removing duplicates and filtering out null values. The dataset contains cleaned data, systematically organized with MP3s, extracted audio features, and album covers for proper storage for subsequent processing. The persistence layer guarantees data integrity and thus forms a robust base for the whole system.

5.3. Business Logic Layer

This includes data preprocessing, model training, and the Recommendation Engine. First, audio features are read by librosa and standardized by a StandardScaler. After that, these features are used as the input to different types of machine learning models: Random Forest, SVM, KNN, XGB, and Gradient Boosting classifiers. Later, their outputs will be stacked in a hybrid model to increase the accuracy in genre classification. It utilizes this layer through a cosine similarity metric measurement that provides the possibility of content-based recommendations by closeness in feature vectors.

It is also at this point that the business logic incorporates feature importance analysis, which helps calculate the contribution of individual features toward the model's predictions. It is through these analyses that the hybrid model learns how to apply only the most important features in the classification of genres and recommendation.

5.4. System Integration

Ultimately, all of this links back to the User Interface Layer, finally hosted by Streamlit, where results are presented to the user with the predicted genre and recommended tracks. Recommendations include track details and album covers, therefore providing a rich and engaging user experience. This seamless integration across layers ensures that the system processes user inputs effectively and returns results in the form of recommendations tailored to their preferences in music.

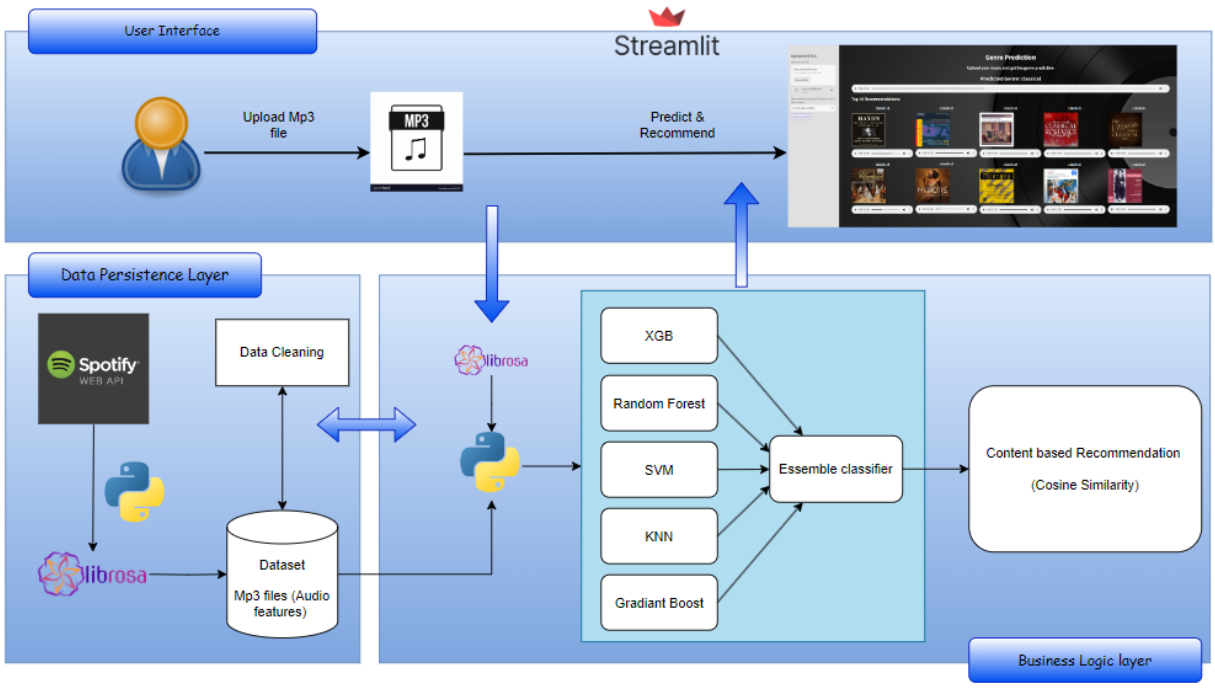

Figure 2: Design Specification

6. Results

6.1. Evaluating Feature Importances with MLP Neural Network Model

The first section focuses on evaluating the feature importances for each genre using a Multi-Layer Perceptron (MLP) Neural Network model. This analysis provides a better understanding of the interrelationships among the audio features within the genres. The feature importances were derived from the MLP Neural Network model, with values normalized between 0 and 1 to indicate the relative importance of each feature.

The x-axes of the bar charts represent the different audio features, such as "mfcc1_mean," "mfcc2_mean," "chroma_stft_mean," etc. The y-axes represent the normalized importance values of these features, ranging between 0 and 1, indicating the relative importance of each feature for classifying the respective music genres.

The results show that Mel-Frequency Cepstral Coefficients (MFCCs) are the most significant features for classifying music genres. MFCCs efficiently describe timbral texture—information crucial in distinguishing genres. Specifically:

- **MFCC1** characterizes overall loudness.
- **MFCC2** and **MFCC3** describe the spectral envelope shape.
- Higher-order MFCCs capture finer details in the spectrum.

Their high importance is due to their close relation to human hearing and their ability to provide a compact, complete, and holistic representation of audio spectral properties. The prominence of MFCCs highlights their effectiveness in capturing essential timbral characteristics that differentiate various music genres.

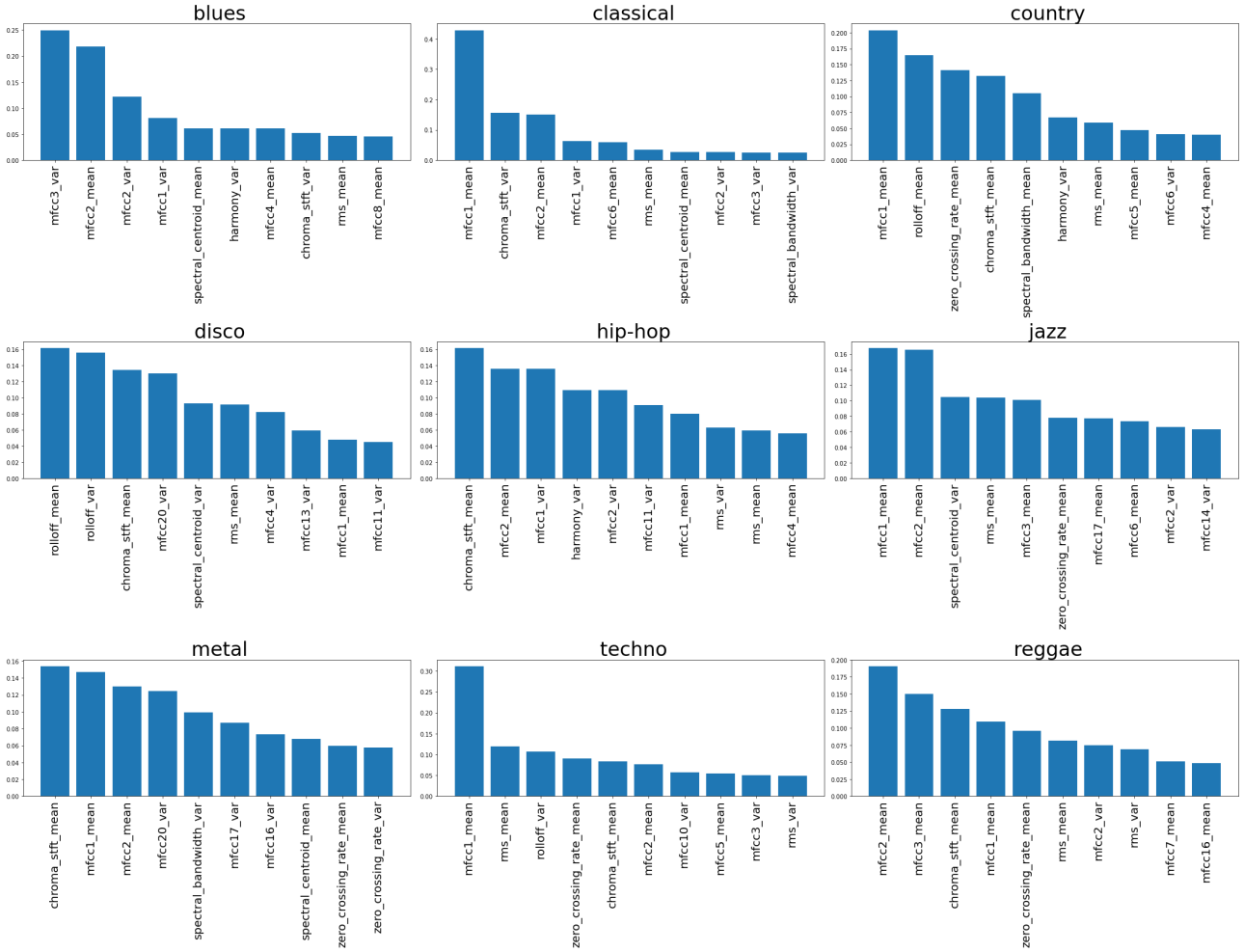

Figure 3: Feature Importances

6.1.1. *Key findings within feature importances*

Based on the feature importance barcharts, we can infer the following: Genres like classical, techno, and disco might be classified with higher accuracy due to their reliance on very specific features. For example:

- **Classical:** Gives a very high importance to MFCC1_mean, indicating that the overall shape of the spectral envelope is very important in its classification.
- **Techno:** MFCC1_mean is also highly important, indicating that specific spectral characteristics are key for identifying this genre.
- **Disco:** Relies heavily on spectral rolloff (both mean and variance) and chroma_stft_mean, suggesting that high-frequency content is a distinguishing factor for this genre.

6.1.2. *Common Features*

Genres such as hip-hop, reggae, jazz, and blues share several important features, indicating a higher potential for misclassification among them. For example:

- **Hip-Hop and Reggae:** Both have chroma_stft_mean and MFCC2_mean as important features, pointing to potential overlap and thus higher chances of misclassification between these two genres.
- **Jazz and Blues:** Show high importance for MFCCs (e.g., MFCC1_mean, MFCC2_mean, MFCC3_var), suggesting they might be challenging to distinguish from each other based solely on these features.

6.1.3. *Feature Overlap*

The presence of common high-importance features across multiple genres suggests areas where classification models might struggle, emphasizing the need for additional or more nuanced features to improve accuracy. For instance:

- **MFCC1_mean:** Appears prominently across multiple genres including classical, country, jazz, metal, techno, and reggae, indicating it captures some common audio characteristics that could lead to misclassification.
- **Chroma_stft_mean:** Important for hip-hop, disco, metal, and reggae, suggesting it captures harmonic content that might be shared among these genres, making them harder to distinguish.

6.2. *Genre Cosine Similarity analysis*

In addition to the feature importance analysis, a Cosine Similarity analysis was done to visualize the similarities between genres based on their audio features calculated using cosine similarity. Figure 4 provides a clear representation of the correlations between genres, indicating potential overlaps and distinctions.

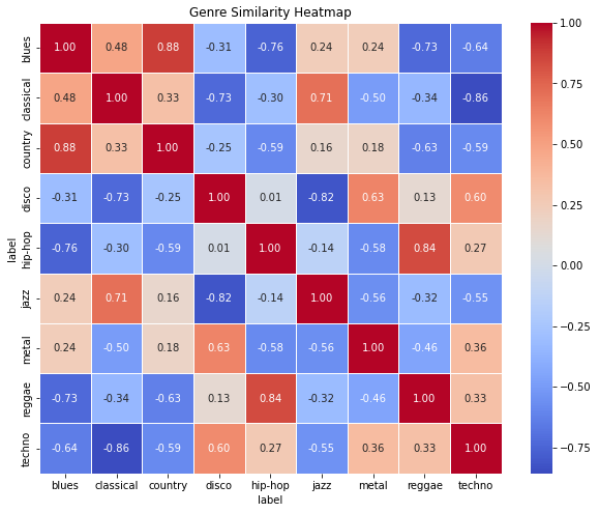

Figure 4: Genre Cosine Similarity

6.2.1. *Observations from the Cosine Similarity Analysis*

The visual representation provided by the in figure 4 displays the cosine similarity scores between different music genres. High similarity is indicated by values closer to 1 (represented by darker red shades), while low similarity is indicated by values closer to -1 (represented by darker blue shades). Scores around 0 (represented by lighter shades) suggest a moderate similarity.

High Similarity Pairs:

- **Blues and Country:** These genres exhibit a high similarity score (0.88), indicating they share many audio features, potentially leading to misclassification between them.
- **Hip-Hop and Reggae:** With a similarity score of 0.84, these genres also share significant audio characteristics, indicating a higher potential for overlap and misclassification.

Low Similarity Pairs:

- **Classical and Techno:** These genres have a low similarity score (-0.82), signifying distinct audio features that aid in accurate classification.
- **Blues and Techno:** Similarly, these genres show a low similarity score (-0.64), suggesting they have distinct audio characteristics, making them easier to differentiate.

General Trends:

- **Distinctive Genres:** Classical and techno genres tend to exhibit lower similarity scores with other genres, reinforcing their reliance on distinctive features for classification.
- **Common Genres:** Genres like blues, jazz, and reggae show higher similarity scores with multiple other genres, highlighting common features that could lead to misclassification.

The Cosine Similarity analysis complements the feature importance analysis by providing a visual representation of genre similarities based on their audio features. It reinforces the observations about distinctive and common features, indicating areas where classification models might perform well and where they might struggle.

6.3. Machine learners & Essembled learner

This analysis was conducted on various machine learning models, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Extreme Gradient Boosting (XGB), Gradient Boosting (GB), and Random Forest (RF). Ultimately, these models were combined into an ensemble classifier. First, the confusion matrix is presented as the outcome of the ensemble classifier, building on the findings from the MLP Neural Network and cosine similarity analysis. Subsequently, the performance metrics of these models will be examined in detail to address the hypotheses.

6.3.1. Confusion Matrix

The confusion matrix for the Meta Model, which combines the outputs of all the evaluated models, is shown below (figure 5). This matrix helps interpret the model's performance across different genres.

- **True Positives (TP)**: Values on the diagonal (e.g., blues correctly predicted as blues - 52).
- **False Positives (FP)**: Off-diagonal values in the predicted genre columns (e.g., predicting blues when it's actually jazz - 27).
- **False Negatives (FN)**: Off-diagonal values in the actual genre rows (e.g., actual blues predicted as another genre - 18).
- **True Negatives (TN)**: All other values not covered by TP, FP, or FN.

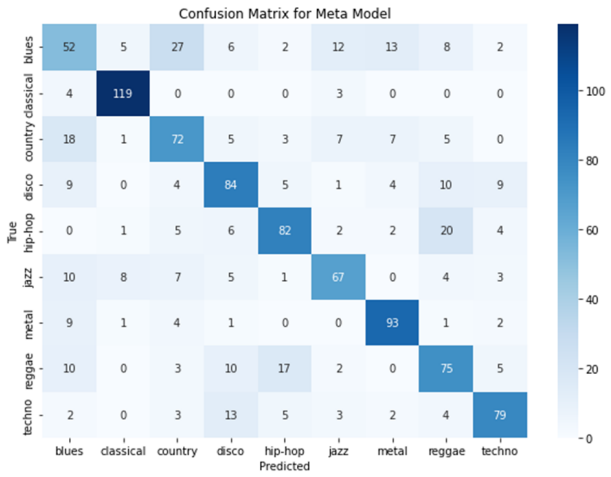

Figure 5: Confusion Matrix

The rows represent the actual genres, and the columns represent the predicted genres. This allows you to see where the model makes correct predictions and where it tends to make mistakes.

High Accuracy Classifications

- **Classical**: Out of 130 instances, 119 were correctly classified. This high accuracy can be attributed to distinctive features like MFCC1 mean and low similarity scores with other genres, such as techno (-0.82).
- **Metal**: With 93 correct classifications, metal benefits from unique features like spectral rolloff and chroma_stft_mean, showing minimal overlap with other genres.

Moderate Accuracy Classifications

- **Jazz**: Achieving 67 correct classifications, jazz's moderate accuracy is due to overlapping MFCC features with genres like blues, as indicated by their high similarity scores.
- **Country**: With 72 correct classifications, misclassifications with blues and reggae are due to shared features and high similarity scores, particularly with blues (0.88).

Low Accuracy Classifications

- **Blues**: Out of 130 instances, only 52 were correctly classified. Misclassifications with country and reggae are significant due to overlapping features and high similarity scores, especially with country (0.88).
- **Reggae**: With 75 correct classifications, reggae often gets misclassified with hip-hop and disco due to shared features and high similarity scores with hip-hop (0.84).

Feature and Similarity Analysis

- **Classical and Techno**: Both genres exhibit distinct characteristics with low similarity scores to other genres (e.g., classical and techno: -0.82), reflected in their high classification accuracy.
- **Hip-Hop and Reggae**: High similarity scores (0.84) explain the frequent misclassifications between these genres, as they share common features.
- **Blues and Country**: High similarity (0.88) leads to significant misclassifications due to shared critical features like MFCCs.

Genres such as classical and metal achieve high accuracy due to their distinctive features and low similarity with other genres. In contrast, genres like blues and reggae demonstrate lower accuracy due to significant feature overlaps with other genres, leading to frequent misclassifications. Understanding these patterns can inform feature selection and model improvements, enhancing the overall performance of music genre classification systems.

6.4. Machine Learning Classification Models Evaluation

To test Hypothesis 1 and Hypothesis 2, the performance of five machine learning models and an ensemble classifier (referred to as the Meta Model) was summarized in Tables 1 and 2. Table 1 utilizes only the MFCC audio features for analysis, while Table 2 uses a combination of all extracted audio features. The performance of these classification models was evaluated using multiple metrics: accuracy, precision, recall, F1 score, and AUC score.

6.4.1. Performance Metrics

Performance metrics such as accuracy, precision, recall, F1 score, and AUC were utilized to evaluate the following models:

- Accuracy: The percentage of correctly classified test samples.
- F-score: The harmonic mean between precision and recall.
- AUC (Area Under the ROC Curve): A measure of the ability of the model to avoid false classification.

Hypothesis 1: Utilizing only Mel-frequency cepstral coefficients (MFCCs) can accurately predict and classify music genres and provide precise genre-based recommendations.

- The performance metrics from Table 1 indicate that using MFCCs alone achieves an accuracy of 0.65 with the Ensemble Classifier, which is higher than the random guessing baseline of 0.50. This suggests that MFCCs are indeed largly effective for music genre classification. Therefore, hypothesis 1 is accepted based on the observed accuracy and performance improvements over random guessing.

Hypothesis 2: The incorporation of timbral, rhythmic, and pitch content features improve the accuracy of automatic music genre classification compared to using MFCC features alone.

- The data from Table 2 demonstrates that including additional audio features (timbral, rhythmic, and pitch content) consistently improves the performance metrics across all models. The Meta Model, in particular, achieves an accuracy of 0.68, further enhancing classification accuracy compared to using MFCCs alone. Therefore, this hypothesis is supported, as the inclusion of these features leads to significant improvements in model performance.

Table 1: Performance of Different Models (MFCCs only)

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC**
Feature Engineering based models					
KNN	0.56	0.56	0.56	0.55	0.85
Random Forest (RF)	0.61	0.61	0.61	0.61	0.90
Gradient Boosting (GB)	0.62	0.61	0.62	0.61	0.90
Support Vector Machines (SVM)	0.64	0.63	0.64	0.63	0.91
XGBoost	0.61	0.61	0.61	0.61	0.91
Ensemble Classifier					
Meta Model	0.65	0.64	0.65	0.64	0.92

Table 2: Performance of Different Models (All Audio Features)

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC**
Feature Engineering based models					
KNN	0.59	0.60	0.59	0.59	0.86
Random Forest (RF)	0.65	0.64	0.65	0.65	0.92
Gradient Boosting (GB)	0.66	0.66	0.66	0.66	0.92
Support Vector Machines (SVM)	0.66	0.66	0.66	0.66	0.93
XGBoost	0.67	0.67	0.67	0.67	0.93
Ensemble Classifier					
Meta Model	0.68	0.67	0.68	0.68	0.93

6.4.2. Data size and consistency

In order to answer Hypotheses 3 and 4, the consistency and size of the dataset were tested. Table 3 uses the GTZAN dataset to test consistency with one individual human classifier, while Table 4 presents the Spotify dataset, classified by multiple human classifiers, using the same number of samples per genre for a valid comparison. To assess the impact of dataset size on accuracy, the Spotify dataset was gradually expanded, and the accuracy rates were visualized in Figure 6.

Data consistency

Hypothesis 3: Using a more consistent dataset (e.g., GTZAN Dataset) results in higher accuracy for music genre classification by machine learning models compared to a dataset classified by multiple people (e.g., a Spotify-extracted dataset).

- The results from both Table 3 (GTZAN Dataset) and Table 4 (Spotify Dataset) show that the models trained on the GTZAN dataset consistently outperform those trained on the Spotify-extracted dataset. Specifically, for the GTZAN dataset, the Meta Model achieves an accuracy of 0.76, whereas the best individual classifier, XGBoost, achieves an accuracy of 0.71. In contrast, for the Spotify dataset, the Meta Model achieves an accuracy of 0.51, while the best individual classifier, Gradient Boosting, achieves an accuracy of 0.56. These results support the hypothesis that a more consistent dataset improves classification accuracy, as evidenced by the higher performance metrics (accuracy, precision, recall, F1-score, and AUC) observed for the GTZAN dataset compared to the Spotify dataset.

Table 3: Gtzan Dataset(100 tracks per genre)

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC**
Feature Engineering based models					
KNN	0.68	0.71	0.70	0.68	0.93
Random Forest (RF)	0.69	0.68	0.71	0.69	0.95
Gradient Boosting (GB)	0.69	0.70	0.70	0.69	0.94
Support Vector Machines (SVM)	0.69	0.70	0.71	0.70	0.96
XGBoost	0.71	0.72	0.72	0.71	0.95
Ensemble Classifier					
Meta Model	0.76	0.75	0.76	0.75	0.97

Table 4: Spotify Dataset (with 100 tracks per genre)

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC**
Feature Engineering based models					
KNN	0.43	0.50	0.46	0.44	0.79
Random Forest (RF)	0.53	0.55	0.56	0.55	0.89
Gradient Boosting (GB)	0.56	0.58	0.58	0.57	0.86
Support Vector Machines (SVM)	0.48	0.49	0.51	0.49	0.88
XGBoost	0.51	0.54	0.54	0.53	0.88
Ensemble Classifier					
Meta Model	0.59	0.60	0.60	0.60	0.90

Data size

Hypothesis 4: Increasing the size of a dataset leads to better genre classification accuracy compared to a smaller dataset.

- The chart in Figure 6, for the Spotify dataset, shows that as the sample size per genre increases, the accuracy of all models improves. For example, the Meta Model's accuracy increases from 0.60 with 100 samples per genre to 0.85 with 600 samples per genre. This demonstrates that larger sample sizes positively affect classification accuracy. However, it is important to note that this holds true only up to a specific amount, as excessively large sample sizes can lead to overfitting, where the model performs well on training data but poorly on unseen data.

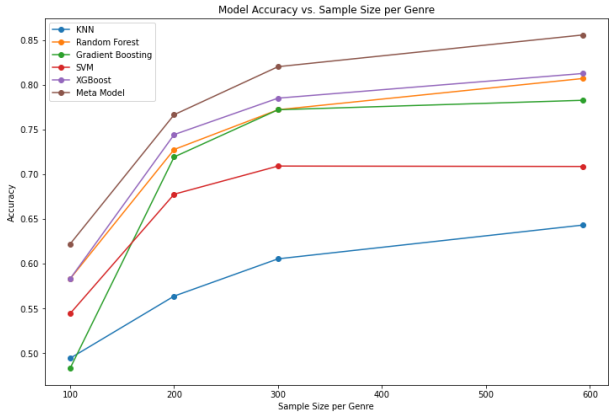

Figure 6: Data Enrichment (Spotify Dataset)

6.4.3. Stacked modeling

To test Hypothesis 5, a comparison was made across all performance metrics between the individual machine learning models and the Meta Model to determine if the Meta Model outperforms the individual models. This comparison aims to distinguish whether the Meta Model improves accuracy.

Hypothesis 5: Utilizing multiple machine learning techniques in a single stacking model will improve the accuracy of genre classifications.

- The results indicate that the Meta Model (stacking model) consistently outperforms individual classifiers across all performance metrics (Tables 1, 2, 3, and 4). These findings support the hypothesis that a stacking model, which integrates multiple machine learning techniques, enhances classification accuracy.

6.5. Recommendation System Evaluation

The recommendation system's performance was evaluated based on genre accuracy and the precision of the recommendations. The evaluation focused on the precision of 10 recommendations per newly added anchor track. To test Hypothesis 6, the full extended table is presented in the appendix, providing an overview of the recommendations made by the system. However, to provide a concise representation, Table 5 is presented as a small visualization of the system's performance.

Genre	Tracks	Correctly Identified	Correctly Predicted	Incorrectly Predicted	Avg Cosine Similarity
Reggae	reggae.00028.wav 10	Yes	5	5	0.603076
Jazz	jazz.00004.wav 10	Yes	8	2	0.764065
Hip-hop	hiphop.00001.wav 10	Yes	6	4	0.618577
Disco	disco.00001.wav 10	Yes	6	4	0.842025
Country	country.00000.wav 10	Yes	5	5	0.740488
Classical	classical.00000.wav 10	Yes	10	0	0.715917
Techno	Clara - techno 10	Yes	10	0	0.634979
Metal	metal.00000.wav 10	Yes	9	1	0.872069
Blues	blues.00000.wav 10	Yes	7	3	0.726119

Table 5: Music Genre Predictions and Accuracy

Hypothesis 6: Music recommendations based solely on audio-level features may provide more diverse recommendations that include tracks from different genres due to the overlap in audio characteristics.

- The results from Table A.6 in the appendix support this hypothesis by showing that anchor tracks often include recommendations from multiple genres. For instance, the reggae track "reggae.00028.wav" has recommendations from hip-hop, techno, and metal, and the jazz track "jazz.00004.wav" includes classical and reggae tracks. This demonstrates that audio-level features can capture similarities across different genres, leading to diverse recommendations.

  However, this diversity can be both beneficial and problematic. While overlapping audio features can make hip-hop and reggae recommendations contextually relevant, mismatches like recommending metal for techno might not align with user preferences. Therefore, while audio-level features promote diversity, the relevance of these recommendations varies, suggesting the need for refining the balance between diversity and genre-specific accuracy.

The genre classification system correctly identified the genre for every song tested, demonstrating high genre accuracy. However, not all recommendations matched the identified genre. This discrepancy is due to the cosine similarity metric, which does not always prioritize songs within the same genre as the highest matches.

This behavior can be attributed to the overlap in audio features across different genres. For instance, genres like reggae and hip-hop or jazz and blues can share similar rhythmic and melodic characteristics, leading the cosine similarity metric to identify them as closely related, even though they belong to different genres. This overlap in audio features, such as spectral properties and rhythmic elements, can cause the system to recommend songs from adjacent genres with high similarity scores.

7. Discussion

7.1. Main Findings

The primary objective of this study was to explore how individual audio features influence the accuracy of music genre classification and recommendation systems, building on existing literature and providing new insights.

7.1.1. Importance of MFCCs

In this research it is confirmed that the Mel-Frequency Cepstral Coefficients (MFCCs) are significant for genre classification. The MFCC1 mean seems to best represent the timbral texture of audio signals, which is key for distinguishing between different genres. This result aligns with the foundational work by Tzanetakis & Cook (2002), who highlighted the importance of timbral texture, rhythmic content, and pitch content for genre classification. This study achieved 65% accuracy using only MFCCs. This shows how powerful the discriminative capability of MFCCs is, suggesting they alone can provide a robust foundation for music genre classification while simplifying feature extraction processes.

7.1.2. Additional Audio Features

Building on this, incorporating additional audio features such as timbral, rhythmic, and pitch content further enhanced classification accuracy. The Meta Model, which integrated these features, achieved an accuracy of 68% on the Spotify dataset. This slight improvement of 3% indicates that while MFCCs are highly significant, combining them with other features can enhance classification performance. For example, the classification of disco benefited significantly from spectral rolloff features, showing that extra features might be necessary for certain genres (e.g. Disco) to achieve optimal classification.

7.1.3. Dataset Consistency

It is also found that the consistency of a dataset plays a crucial role in classification accuracy. West & Cox (2004) discussed how cultural and linguistic factors influence genre definition and classification. The study demonstrated that models trained on the more consistent GTZAN dataset (classified by a single individual) outperformed those trained on a Spotify-extracted dataset (classified by multiple individuals). The Meta Model achieved 76% accuracy with the GTZAN dataset compared to 51% with the Spotify dataset (same amount of samples per genre). This finding highlights how cultural and linguistic factors in human classification can affect the overlap of audio features in different genres.

Cultural differences and the inherent subjectivity of genre definitions lead to inconsistencies in human classification on platforms like Spotify. Genres often overlap in audio features and are not always clearly defined, which can result in varied human classifications. These overlaps are evident when you look at the misclassification rates of the models. Genres such as hip-hop and reggae or jazz and blues show higher misclassification rates because there is also higher similarity between their audio-level features. While some of these features are very discriminative, others may add noise and hence mislead genre prediction. Building on the work of Lippens et al. (2004), it has been reported that human genre classification is subjective and difficult, reflected in the complexity of obtaining very high accuracy with automatic systems. This cultural and feature-related ambiguity in genre classification is mirrored in both the feature importances and misclassification patterns as shown in the findings.

7.1.4. Dataset Size

Considering the unreliability of human classification, it is essential to investigate whether increasing dataset size can enhance model performance. As sample size per genre increased, accuracy improved for all models. For instance, the Meta Model's accuracy rose from 60% with 100 samples per genre to 85% with 600 samples. This indicates that larger datasets provide more comprehensive training data, enabling models to better detect underlying patterns and nuances within the music, including sub-genres. However, this trend holds only to a certain extent, as excessively large datasets may cause overfitting (Kavzoglu, 2009), resulting in models that perform well on training data but poorly on new, unseen data.

7.1.5. Ensemble Learning Techniques

Contributing to the finding of Li et al. (2003) and Ahlem et al. (2022) that a combination of different sets of features and machine learning techniques gives great enhancement to these models since they will be robust and provide deeper insights through multiple models. Specifically, ensemble learning in the form of the stacking model was found to be critical for improving classification accuracy. In most cases, the Meta Model was better than any dimensionality-reduced single classifier with respect to both accuracy and AUC scores. Ensemble methods are, therefore, very effective in the classification of music genres due to their capability of leveraging various algorithms' strengths for more accurate and robust predictions.

7.1.6. Content based Recommendations

Lastly, the study found that recommendations based solely on the cosine similarity of audio features can vary in accuracy depending on the distinctiveness of the genre's audio features. Genres whose audio is characterized by more unique and pronounced features, such as classical and metal, recommend similar genres more accurately and consistently. However, for genres with less distinctive audio features and more overlap with other genres, the accuracy of recommendations decreases. For instance, an anchor track from a genre like blues, which shares many audio features with jazz, might lead to less accurate recommendations if you only look at the genre classification as a threshold.

It is important to note that whilst the genre might be misclassified, the recommendation might very well speak to the users preference. For instance, a blues song recommended as jazz might be wrongly classified, since these genres share similar aspects, but could be a preferable recommendation based on the anchor song. Therefore, even if the recommendation may not exactly match the original genre, contextually it may be relevant and appropriate. This nuance thus underlines the complexity of the relations between genres, demanding further research in classification and recommendation systems, taking into account both the distinctiveness and overlap of audio features across genres.

Overall, this study contributes valuable insights to the field of Music Information Retrieval (MIR) by demonstrating the critical role of MFCCs, the benefits of a multi-feature approach, the importance of dataset consistency and size, and the effectiveness of ensemble learning techniques. Taking these factors into consideration, it can be assured that future research in music genre classification and recommendation will enhance accuracy and reliability significantly and improve the user experience in music streaming applications.

7.2. Scientific Contribution

This paper contributes to the area of MIR, since it specifically determines the role of Mel-Frequency Cepstral Coefficients in genre classification. Building on the foundational work of Tzanetakis & Cook (2002), the study confirms that MFCCs effectively capture the timbral texture of audio signals, achieving high classification accuracy, also using a dataset that is less consistent of predefined genre classification. Furthermore, incorporating additional features like timbral, rhythmic, and pitch content further enhances accuracy, demonstrating the value of a multi-feature approach.

The study also highlights the effectiveness of ensemble learning techniques, particularly stacking models, in improving classification performance. By combining multiple machine learning methods, the Meta Model achieved superior accuracy, reinforcing findings by Li et al. (2003) and Ahlem et al. (2022). Additionally, the content-based recommendation system developed using cosine similarity addresses the cold-start problem, offering accurate and relevant music recommendations based on audio features.

7.3. Managerial Relevance

The automated genre classification models developed in this study can help music distributors and platforms streamline their operations. These models reduce the need for manual classification, thereby increasing operational efficiency and accuracy in genre tagging.

Moreover, it may increase the rate of user retention and subscription with the help of enhanced music recommendation. The more personalized and precise music recommendations streaming services provide, the more satisfied their users will be, which ultimately benefits their business models.

7.4. Limitations and Recommendations

Though this study yields many valuable insights, there are several limitations that suggest the need for further investigation in several areas. The following points outline the key areas for potential enhancement and future research:

1. **Dataset Enrichment**: It was found that increasing the dataset during the research led to improved model performance. A larger dataset provides more values to train on, thus enriching the dataset could further increase the accuracy of the models. Future research should consider expanding the dataset to include a more extensive and diverse collection of music tracks, genres, and sub-genres.
2. **Inclusion of Advanced Machine Learning Models**: In the study, it was indicated that by integrating more machine learning models, the deep learning models such as Neural Networks (NN) would bring many benefits. According to the theoretical background, Convolutional Neural Networks had been proved to have high accuracy. Although CNNs were considered "black-box" models and could be less interpretable because of their complexity, they could provide valuable results. Future studies may integrate such advanced models in order to achieve better classification accuracy.
3. **Integration of User Experience Metrics**: The findings of this study have several significant implications for the music industry and related fields. Currently, a user experience metric is not part of the recommendation system in the study, so it is hard to determine whether or not the recommended system of the tracks will really fit the preferences of the users. Enhancing audio platform systems by integrating a metric to measure user satisfaction based on music audio features could significantly improve recommendation accuracy and user engagement.
4. **Increasing Variety of Audio Features**: The findings indicated that utilizing a broader range of audio features improves model accuracy. Therefore, it is recommended for future research to increase the number of audio features considered. This would help in fully describing audio signals and hence improved genre classification results and recommendations.
5. **Differentiation of Intrinsic Genre Relationships**: This paper identified intrinsic relationships between certain genres, which share similar importance in audio features and exhibit significant similarities. Understanding these relationships is crucial for improving model accuracy. Future research should focus on developing methods to differentiate these similarities effectively. This could involve investigating new features or refining existing ones to better distinguish between genres with overlapping characteristics.

7.5. Conclusion

In conclusion, this study has demonstrated the significant impact of individual audio features on the accuracy of music genre classification and recommendation systems. The findings highlight the critical role of MFCCs and the benefits of a comprehensive feature set and ensemble learning techniques. The practical implications for the music industry are substantial, offering pathways for improved user engagement and operational efficiency. Future research should aim to address the identified limitations and continue to explore innovative approaches to enhance automatic genre classification and personalized music recommendations.

References

Abeysinghe, A., Fard, M., Jazar, R., Zambetta, F., Davy, J., 2021 . Mel frequency cepstral coefficient temporal feature integration for classifying squeak and rattle noise. The Journal of the Acoustical Society of America 150 (1), 193–201.

Afchar, D., 2023 . Interpretable music recommender systems. PhD Thesis, Sorbonne Université.

Ahlem, D. R. I. F., Saadeddine, S., Hocine, C., 2022 . An interactive attention network with stacked ensemble machine learning models for recommendations. In: Optimization and Machine Learning: Optimization for Machine Learning and Machine Learning for Optimization, pp. 119–150.

Aljanaki, A., Soleymani, M., 2018 . Mid-level representations for music information retrieval.

Amit, Y., Geman, D., 1997 . Shape quantization and recognition with randomized trees. Neural Computation 9 (7), 1545–1588.

Breiman, L., 1996 . Bagging predictors. Machine Learning 24, 123–140.

Burred, J. J., Lerch, A., 2004 . Hierarchical automatic audio signal classification. Journal of the Audio Engineering Society 52 (7/8), 724–739.

Chapman, P., Clinton, J. M., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. R., Wirth, R., 2000 . Crisp-dm 1.0: Step-by-step data mining guide.

Chathuranga, D., Jayaratne, L., 2013 . Automatic music genre classification of audio signals with machine learning approaches. GSTF Journal on Computing (JoC) 3, 1–12.

Cortes, C., Vapnik, V., 1995 . Support-vector networks. Machine Learning 20, 273–297.

Darshna, P., 2018 . Music recommendation based on content and collaborative approach & reducing cold start problem. In: 2018 2nd International Conference on Inventive Systems and Control (ICISC), pp. 1033–1037.

Downie, J. S., 2003 . Music information retrieval. Annual Review of Information Science and Technology 37 (1), 295–340.

Elbir, A., Aydin, N., 2020 . Music genre classification and music recommendation by using deep learning. Electronics Letters 56 (12), 627–629.

Ellis, D., 2007 . Chroma feature analysis and synthesis. `http://labrosa.ee.columbia.edu/projects/chroma`.

Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K., 2003 . Knn model-based approach in classification. In: On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings. Springer Berlin Heidelberg, pp. 986–996.

Hastie, T., Tibshirani, R., Friedman, J. H., 2009 . The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Vol. 2.

Heikkinen, V., 2012 . Linguistic genre analyses, intertextuality, and ideology. In: Genre–Text–Interpretation, p. 60.

Jiang, D. N., Lu, L., Zhang, H. J., Tao, J. H., Cai, L. H., 2002 . Music type classification by spectral contrast feature. In:

Proceedings. IEEE International Conference on Multimedia and Expo. Vol. 1. IEEE, pp. 113–116.

Kavzoglu, T., 2009 . Increasing the accuracy of neural network classification using refined training data. Environmental Modelling & Software 24 (7), 850–858.

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., Turnbull, D., 2020 . Music emotion recognition: A state of the art review.

Knees, P., Schedl, M., 2013 . A survey of music similarity and recommendation from music context data. ACM Transactions on Multimedia Computing, Communications, and Applications 10 (1), 2:1–2:21.

Knees, P., Schedl, M., 2016 . Music similarity and retrieval: An introduction to audio- and web-based strategies.

Li, T., Ogihara, M., Li, Q., 2003 . A comparative study on content-based music genre classification. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, 282–289.

Li, Z., Peng, H., Tan, S., Zhu, F., 2023 . Music classification with convolutional and artificial neural network 2580 (1), 012059.

Lin, Y. P., Wang, C. H., Wu, T. L., Jeng, S. K., Chen, J. H., 2007 . Multilayer perceptron for eeg signal classification during listening to emotional music. In: TENCON 2007-2007 IEEE Region 10 Conference, pp. 1–3.

Lippens, S., Martens, J.-P., De Mulder, T., 2004 . A comparison of human and automatic musical genre classification. 2004 IEEE international conference on acoustics, speech, and signal processing 4, iv–iv.

Martínez Ramírez, M. A., Benetos, E., Reiss, J. D., 2020 . Deep learning for black-box modeling of audio effects. Applied Sciences 10 (2), 638.

McKay, C., Fujinaga, I., 2009 . jmir: Tools for automatic music classification. ICMC.

Melchiorre, G., Schedl, M., 2020 . High-level audio features for emotion and genre classification.

Mitrović, D., Zeppelzauer, M., Breiteneder, C., 2010 . Features for content-based audio retrieval. In: Advances in Computers. Vol. 78, pp. 71–150.

Muraina, I. O., 2022 . Ideal dataset splitting ratios in machine learning algorithms: general concerns for data scientists and data analysts. In: 7th International Mardin Artuklu Scientific Research Conference. pp. 496–504.

Müller, M., 2015 . Fundamentals of music processing: Audio, analysis, algorithms, applications.

Newsroom, S., 2023 . Spotify debuts a new ai dj right in your pocket.

Pampalk, E., Rauber, A., Merkl, D., 2002 . Content-based organization and visualization of music archives. Proceedings of the tenth ACM International Conference on Multimedia, 570–579.

Rebelo, A., Fujinaga, I., Paszkiewicz, F., Marcal, A., Guedes, C., Cardoso, J., 2012 . Optical music recognition: state-of-the-art and open issues. International Journal of Multimedia Information Retrieval 1 (3), 173–190.

Sheikh Fathollahi, M., Razzazi, F., 2021 . Music similarity measurement and recommendation system using convolutional neural networks. International Journal of Multimedia Information Retrieval 10, 43–53.

Song, Y., Dixon, S., Pearce, M., 2012 . A survey of music recommendation systems and future perspectives. 9th international symposium on computer music modeling and retrieval 4, 395–410.

Tjoa, S., 2017 . Music information retrieval. `https://musicinformationretrieval.com/spectral_features.html`.

Tzanetakis, G., Cook, P., 2002 . Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing 10 (5), 293–298.

West, K., Cox, S., 2004 . Features and classifiers for the automatic classification of musical audio signals. ISMIR.

Wold, E., Blum, T., Keislar, J., Wheaten, D., 1996 . Content-based classification, search, and retrieval of audio. IEEE Transactions on Multimedia 3 (3), 27–36.

Appendix A. Appendix

Table A.6: Music Genre Predictions and Accuracy

Genre	Tracks	Correctly Identified	Correctly Predicted	Cosine Similarity
Reggae	reggae.00028.wav	Yes		
	Kingston dub - reggae		Yes	0.772414
	Hard Times - reggae		Yes	0.748922
	We Nah Leave - reggae		Yes	0.629309
	Rhymes for the Deaf, Dumb and Blind - hip-hop		No	0.628700
	Triangle - techno		No	0.615832
	Paymaster - reggae		Yes	0.599922
	Chalice In Mind - reggae		Yes	0.596985
	Live to Love Dub - reggae		Yes	0.591798
	Mine Is No Disgrace - metal		No	0.587914
	Stranger Danger - hip-hop		No	0.584981
Jazz	jazz.00004.wav	Yes		
	Distinguished Mindbeing Miracles - jazz		Yes	0.842872
	Partite diverse sopra il corale Christ, der du bist der helle Tag - classical		No	0.818312
	Cello Concerto No. 1 in A Minor, Op. 38, R. 190 - classical		No	0.817228
	Bgm for Cafe Drinking - jazz		Yes	0.797353
	Background for Infinite Hours - jazz		Yes	0.792143
	Rolling Home - jazz		Yes	0.789307
	Wonderful Ambience for Night Spirits - jazz		Yes	0.787906
	Orlando Sleepeth, P. 61 - classical		No	0.787389
	Partite diverse super Christ, der du bist der helle Tag - classical		No	0.773858
	Let Him Go - reggae		No	0.763233
Hip-hop	hiphop.00001.wav	Yes		
	Banlieue - hip-hop		Yes	0.708499
	Triumph (feat. Ol' Dirty Bastard, Inspectah Deck) - hip-hop		Yes	0.692490
	More Than a Man - hip-hop		Yes	0.636384
	Kickback (feat. Trello Tha God) - hip-hop		Yes	0.601111
	I LIKE ME BETTER (TECHNO) - techno		No	0.585308
	Broken from the Start - hip-hop		Yes	0.583842
	MAH BUR - hip-hop		Yes	0.573386
	Do What I Gotta Do - hip-hop		Yes	0.564749
	True Believer in Dub - Dub - reggae		No	0.560488
	One Verse - hip-hop		Yes	0.560006
Disco	disco.00001.wav	Yes		
	Goldmine - disco		Yes	0.882473
	Dance pra Valer - reggae		Yes	0.875924
	Crazy Family - Single Mix - disco		Yes	0.875778
	Don't Wait Too Long - disco		Yes	0.850240
	Wer hat an der Uhr gedreht - disco		Yes	0.845825
	Tanze Samba mit mir - Clubmix - disco		Yes	0.844038
	God Love Is All Around - reggae		Yes	0.842067
	Starting All Over - metal		No	0.842036
	Beyond Death - metal		No	0.840519
	Voulva - Alberto Ruiz Remix - techno		No	0.840172

Continued on next page

Table A.6: Music Genre Predictions and Accuracy (continued)

Genre	Tracks	Correctly Identified	Correctly Predicted	Cosine Similarity
Country	country.00000.wav	Yes		
	Um die ganze Welt - country		Yes	0.831314
	Critic - country		Yes	0.767512
	Sunshine Electric Raindrops (Naked) - metal		No	0.750655
	Et Filii - metal		No	0.748034
	Yours for a Song - blues		Yes	0.738459
	Women Ain't Whiskey - country		Yes	0.730308
	Embrace - Madeaux Remix - disco		Yes	0.726954
	In My Mind - disco		Yes	0.725978
	Joe Citizen Blues - Live - country		No	0.725584
	Susurro Nocturno - metal		No	0.724886
Classical	classical.00000.wav	Yes		
	Medea (Medee) (Sung in Italian) Act I March and Chorus - classical		Yes	0.743429
	Der Rosenkavalier, Op. 59, TrV 227 Waltz - classical		Yes	0.726940
	Concerto for Two Oboes, Strings and Basso Continuo - classical		Yes	0.719292
	Canon In D Major - classical		Yes	0.712239
	Phantasy Concerto for Violin and Orchestra, Op. 42 - classical		Yes	0.711845
	Symphony No. 97 in C Major, Hob. 1.97 II. Adagio ma non troppo - classical		Yes	0.706740
	Canon and Gigue in D Major - classical		Yes	0.700660
	Paganini Sonata for Violin and Guitar in D Major - classical		Yes	0.693946
	Allabreve in D - Dur - classical		Yes	0.676999
	Mass in B Minor, BWV 232 Credo. Et exspecto resurrectionem mortuorum - classical		Yes	0.676998
Techno	Clara - techno	Yes		
	Best Memories, Heaviest Tears - metal		No	0.662959
	No Police - Original Mix - techno		Yes	0.648746
	Dark Jazz Dealer - jazz		Yes	0.645680
	Mortal Combat - techno		Yes	0.642902
	Seeing Someone Else - country		No	0.634579
	U Know - Extended Mix - techno		Yes	0.629994
	Der aus dem Jungel kam - techno		Yes	0.624196
	90's Guetto - Original Mix - techno		Yes	0.622687
	Orbital Darkness - Original Mix - techno		Yes	0.622609
	Babamba - techno		Yes	0.622453
Metal	metal.00000.wav	Yes		
	My Killers - blues		Yes	0.896708
	Keep On Rockin - metal		Yes	0.889428
	Peppermint Tribe - metal		Yes	0.886824
	Harder Than Ever - metal		Yes	0.886251
	Morbid Shape in Black - Rough Mix - metal		Yes	0.878505
	Trance - metal		Yes	0.877054
	Odpowiedz I'm - metal		Yes	0.875482
	Let's Shake It Up - blues		Yes	0.874086
	Montag - metal		Yes	0.872719
	Death Favours the Enemy - Live - techno		No	0.849829

Continued on next page

Table A.6: Music Genre Predictions and Accuracy (continued)

Genre	Tracks	Correctly Identified	Correctly Predicted	Cosine Similarity
Blues	blues.00000.wav	Yes	Yes	0.777525
	Blood of the Sun - Live at New Years 1971 - blues		Yes	0.747305
	Six Days on the Road (Live) - country		Yes	0.738899
	Straight Arrow - country		Yes	0.721388
	What a Little Moonlight Can Do (with Teddy Wilson & His Orchestra) - blues		Yes	0.688027
	Goodbye Yellow Brick Road - country		No	0.677679
	Stranger On The Shore - reggae		Yes	0.672251
	Bad Indian - blues		Yes	0.666168
	Nocturnes, L. 91 II. Fêtes - classical		Yes	0.656532
	T.S.O.P. (The Sound Of Philadelphia) (feat. The Three Degrees) - disco		Yes	0.649115
	Symphony No. 9 in D Major IV. Adagio - Sehr langsam und noch zurckhaltend - classical			