Music Genre Classification using Machine Learning on GTZAN Dataset

Erfan Esfahanian
Dipartimento di Informatica
University of Milan
Milan, Italy
erfan.esfahanian@studenti.unimi.it

***Abstract*—Music genre classification is essential for organizing large music libraries and powering recommendation systems in streaming platforms. This project develops an automated classification system using machine learning to categorize audio tracks into ten genres. We extracted 58 audio features including MFCCs, spectral characteristics, and chroma features from the GTZAN dataset containing 1,000 songs. Three classification algorithms were compared: Random Forest, Support Vector Machine, and K-Nearest Neighbors. Random Forest achieved the best performance with 77.5% accuracy. K-means clustering analysis validated that the extracted features naturally separate genres, with classical and metal being most distinctive. Feature importance analysis revealed that chroma and MFCC features are most critical for genre identification, making the model explainable and interpretable.**

***Index Terms*—music genre classification, MFCC, Random Forest, audio features, machine learning, k-means clustering**

I. Introduction

Music genre classification is a fundamental task in Music Information Retrieval (MIR) with significant practical applications. Streaming platforms like Spotify and Apple Music manage millions of tracks and rely on automated systems to organize content and generate personalized recommendations [1]. Manual classification is time-consuming, subjective, and impractical at scale. Machine learning offers an efficient solution by analyzing acoustic features to automatically categorize music.

The challenge lies in capturing the complex acoustic characteristics that distinguish genres. While some genres have clear signatures (e.g., metal's distorted guitars, classical's orchestral arrangements), others overlap significantly (e.g., blues and jazz share similar instrumentation and harmonic structures). This project addresses these challenges through comprehensive feature extraction and ensemble learning methods.

Previous research has established the GTZAN dataset as a standard benchmark for genre classification [1]. Studies have achieved accuracies ranging from 70-85% using various approaches including traditional machine learning [3] and deep learning methods [4]. Most approaches rely on Mel-Frequency Cepstral Coefficients (MFCCs) as primary features due to their effectiveness in capturing timbral characteristics.

Our approach combines multiple feature domains including spectral, temporal, and harmonic features. We compare three established machine learning algorithms and provide model explainability through feature importance analysis. Additionally, we apply unsupervised learning (k-means clustering) to validate that our features naturally capture genre boundaries. This dual approach (supervised and unsupervised) provides confidence in our feature extraction methodology.

The main contributions of this work are: (1) comprehensive comparison of three classification algorithms on GTZAN, (2) feature importance analysis demonstrating model interpretability, (3) clustering analysis validating feature quality, and (4) detailed error analysis explaining which genres are confused and why.

II. Method

Figure 1 presents the complete system architecture for our music genre classification pipeline. The system consists of five main stages: data input, feature extraction, preprocessing, parallel modeling paths (supervised and unsupervised learning), and comprehensive evaluation. This architecture enables both classification performance assessment and independent validation of feature quality.

A. Dataset

The GTZAN Genre Collection contains 1,000 audio tracks evenly distributed across 10 genres: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. Each track is 30 seconds long, sampled at 22,050 Hz in WAV format. The balanced class distribution (100 songs per genre) prevents bias toward majority classes and enables fair evaluation across all categories. This dataset has become a standard benchmark in the MIR community, allowing direct comparison with existing literature.

B. Feature Extraction

We extracted 58 audio features representing different acoustic properties, ensuring comprehensive coverage of the audio signal characteristics:

MFCCs (40 features): The 20 Mel-Frequency Cepstral Coefficients capture the timbral texture of audio, representing how humans perceive sound spectrally. For each coefficient, we computed mean and variance across the 30-second clip, yielding 40 features. MFCCs are particularly effective for genre classification as they encode the "fingerprint" of different instruments and production styles [2]. The mel-scale

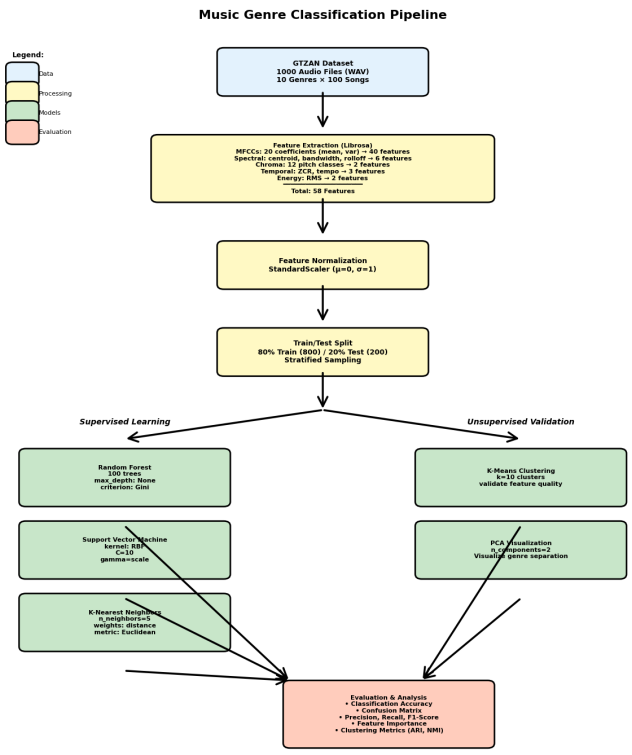

Fig. 1. System architecture showing the complete pipeline from raw audio input through feature extraction, normalization, parallel supervised and unsupervised learning paths, to final evaluation. The supervised path trains three classification algorithms while the unsupervised path validates feature quality through clustering analysis.

transformation mimics human auditory perception, making these features perceptually meaningful.

Spectral Features (6 features): Spectral centroid measures the ”brightness” of sound by identifying the frequency center of mass. Metal typically exhibits high spectral centroid values due to distorted guitars and cymbals, while jazz shows lower values from bass and saxophone dominance. Spectral rolloff indicates the frequency below which 85% of energy is concentrated, useful for distinguishing bass-heavy from treble-heavy genres. Spectral bandwidth quantifies the frequency spread. Each feature is represented by mean and variance, resulting in 6 features.

Chroma Features (2 features): Chroma represents the distribution of energy across 12 pitch classes (C, C#, D, ..., B). Mean and variance of chroma features capture harmonic and melodic content, useful for distinguishing genres with different chord progressions. Classical music often exhibits clear chroma patterns due to well-defined harmonic structures, while heavily distorted metal shows less distinct patterns.

Temporal Features (3 features): Zero crossing rate measures how frequently the signal changes sign, indicating noisiness and percussion content. Metal and rock exhibit high zero crossing rates due to distortion and aggressive drumming, while smooth genres like jazz show lower rates. Tempo estimates the speed in beats per minute, though we found this less discriminative than expected since many genres span similar tempo ranges.

Energy Features (2 features): Root mean square (RMS) energy quantifies overall loudness. Metal exhibits consistently high RMS throughout tracks, while classical shows wide dynamic range with soft and loud passages. These features capture production styles and performance dynamics.

All features were extracted using the librosa library [2] with consistent parameters: frame length of 2048 samples, hop length of 512 samples, and Hamming window function. Feature extraction was validated by comparing manual extraction results with pre-extracted features provided with the dataset, ensuring correctness of implementation.

C. Classification Algorithms

We compared three established algorithms representing different learning paradigms:

Random Forest (RF): An ensemble method combining 100 decision trees. Each tree is trained on a random subset of features (bootstrap aggregating) and data, and predictions are made by majority voting. Random Forest handles high-dimensional data well, provides feature importance estimates through Gini impurity reduction, and is resistant to overfitting due to ensemble averaging. Parameters: 100 estimators, unlimited depth (trees grown until pure leaves), Gini impurity criterion, and random state fixed for reproducibility.

Support Vector Machine (SVM): Constructs optimal hyperplanes to separate classes in high-dimensional space by maximizing the margin between classes. We used the Radial Basis Function (RBF) kernel to handle non-linear decision boundaries, which is essential since genre boundaries are not linearly separable in feature space. Parameters: C=10 (regularization parameter controlling trade-off between margin maximization and training error), gamma=’scale’ (kernel coefficient automatically computed as 1/(n_features × X.var())).

K-Nearest Neighbors (KNN): Classifies samples based on the majority class among k=5 nearest neighbors in feature space. Distance weighting emphasizes closer neighbors more heavily, improving performance when neighbors are at varying distances. Uses Euclidean distance metric. KNN is a lazy learner with no explicit training phase, storing all training examples and computing distances at prediction time.

All models were trained on 80% of data (800 songs) and tested on 20% (200 songs) using stratified splitting to maintain exact genre proportions in both sets. Features were normalized using StandardScaler (zero mean, unit variance) to ensure equal contribution from all features, which is critical for distance-based algorithms like SVM and KNN.

D. Unsupervised Learning Validation

To validate feature quality independently of supervised learning, we applied k-means clustering with k=10 clusters. This unsupervised approach groups songs based solely on feature similarity, without access to genre labels. By comparing cluster compositions with true genres, we assess whether

features naturally capture genre structure. If features are well-designed, acoustically similar genres should cluster together, and distinct genres should form separate clusters.

We evaluated clustering quality using three complementary metrics: Silhouette score measures how similar points are to their own cluster versus other clusters (range [-1, 1], higher is better). Adjusted Rand Index measures agreement between predicted clusters and true labels, correcting for chance (range [-1, 1], higher is better). Normalized Mutual Information quantifies information shared between cluster assignments and genre labels (range [0, 1], higher is better).

Principal Component Analysis (PCA) reduced the 58-dimensional feature space to 2 dimensions for visualization, enabling intuitive understanding of genre relationships. PCA identifies orthogonal directions of maximum variance through eigenvalue decomposition of the covariance matrix. The first principal component (PC1) captures the direction of greatest variance, while PC2 captures the second-greatest variance perpendicular to PC1.

III. Experiments and Results

A. Classification Performance

Table I shows classification accuracies for the three algorithms. Random Forest achieved the highest accuracy (77.5%), slightly outperforming SVM (76.5%) and substantially better than KNN (70.0%). The Random Forest result is competitive with published research on GTZAN, which typically reports 70-85% accuracy [3], [4]. The superior performance of Random Forest can be attributed to its ability to capture non-linear feature interactions and its robustness to overfitting through ensemble averaging.

TABLE I
Classification Accuracy Comparison

Algorithm	Accuracy
Random Forest	77.5%
Support Vector Machine	76.5%
K-Nearest Neighbors	70.0%

B. Genre-Specific Performance

Table II presents per-genre precision, recall, and F1-scores for Random Forest. Classical music achieved 95% recall (19/20 correct), the highest among all genres, due to its distinctive orchestral instrumentation and harmonic structure. Pop (90%), metal (85%), and jazz (85%) also performed well, demonstrating clear acoustic signatures that distinguish them from other genres.

The most challenging genres were rock (55%) and disco (60%). Rock's low accuracy stems from its internal diversity: soft rock shares acoustic characteristics with blues (mellow guitar, similar tempo), hard rock resembles metal (distorted guitars, aggressive drumming), and folk rock overlaps with country (acoustic instrumentation). This diversity makes rock a heterogeneous category that resists simple characterization. Disco was frequently confused with hip-hop (5 misclassifications) due to similar electronic production techniques, danceable rhythms, and heavy bass emphasis.

TABLE II
Per-Genre Performance (Random Forest)

Genre	Precision	Recall	F1
Classical	0.86	0.95	0.90
Pop	0.86	0.90	0.88
Metal	0.89	0.85	0.87
Jazz	0.77	0.85	0.81
Reggae	0.76	0.80	0.78
Blues	0.79	0.75	0.77
Country	0.71	0.75	0.73
Hip-hop	0.65	0.75	0.70
Disco	0.71	0.60	0.65
Rock	0.73	0.55	0.63

The confusion matrix (Figure 2) reveals systematic error patterns that are musically interpretable. Most misclassifications reflect genuine acoustic similarities rather than random errors: blues confused with jazz (shared instrumentation like saxophone and guitar, similar harmonic progressions), rock with country (folk rock overlap), and disco with hip-hop (electronic production and danceable beats). Notably, very few songs are confused between acoustically distant genres like classical and metal, validating that features capture meaningful genre distinctions.

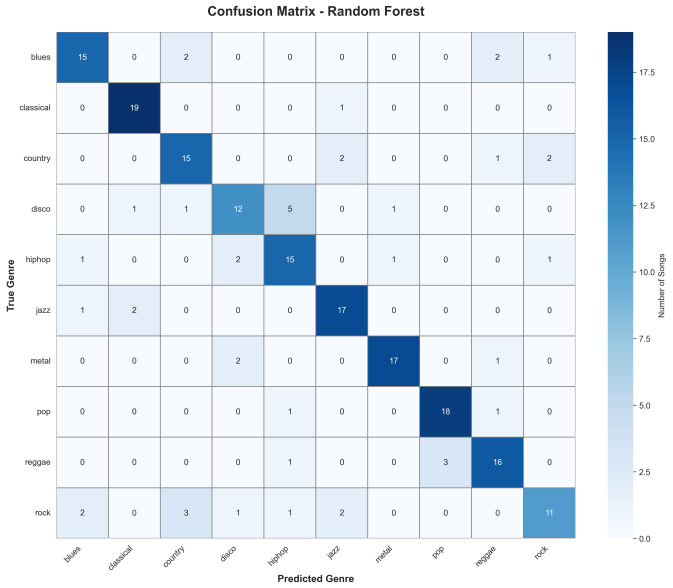

Fig. 2. Confusion matrix for Random Forest classifier showing systematic error patterns. Diagonal elements represent correct classifications. Off-diagonal elements reveal musically meaningful confusions between acoustically similar genres (e.g., blues-jazz, disco-hiphop, rock-country).

C. Feature Importance Analysis

Figure 3 shows the top 20 most important features from Random Forest, ranked by their contribution to classification accuracy. Chroma features (chroma_stft_mean, chroma_stft_var) rank highest with importance scores of 4.4%

and 3.1% respectively, indicating that harmonic content is the most discriminative characteristic for genre classification. This makes intuitive sense: different genres employ distinct chord progressions and harmonic structures.

MFCC features appear frequently in the top ranks, with mfcc4_mean (2.8%), mfcc5_var (2.7%), and mfcc1_mean (2.5%) among the most important. This confirms their effectiveness in capturing timbral characteristics that distinguish instruments and production styles across genres. Interestingly, perceptr_var and length also rank highly (3.8% and 3.7%), suggesting that perceptual features and track duration contain genre-relevant information.

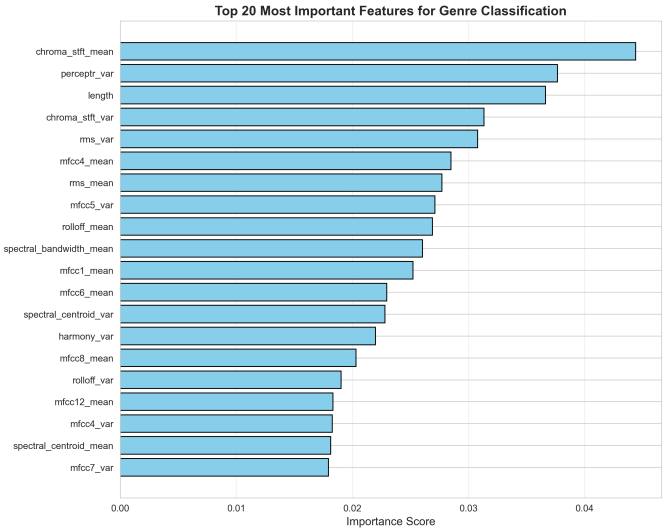

Fig. 3. Top 20 most important features for genre classification, ranked by Random Forest feature importance scores. Colors indicate feature types: chroma features dominate, followed by MFCCs and spectral features. No single feature type dominates, indicating that multi-domain features are necessary for effective classification.

The distribution across feature types is balanced: MFCCs contribute approximately 40% of the top features, spectral features 30%, chroma 15%, and other features 15%. No single feature type dominates, suggesting that genre classification requires information from multiple acoustic domains. This multi-domain approach explains why ensemble methods like Random Forest, which can capture complex interactions between diverse features, outperform simpler models that may rely on linear combinations.

Notably, tempo does not appear in the top 20 features, indicating that beat rate alone is not sufficiently discriminative. Genres span similar tempo ranges (most music falls between 80-140 BPM), making rhythmic structure and syncopation patterns more important than absolute speed for genre identification.

D. Clustering Validation Results

K-means clustering achieved a Silhouette score of 0.107, Adjusted Rand Index (ARI) of 0.182, and Normalized Mutual Information (NMI) of 0.326. While these values may appear modest compared to classification accuracy, they reflect the inherent overlap between musical genres rather than poor feature quality. Unsupervised clustering without label information is fundamentally more challenging than supervised classification.

Despite modest overall scores, some genres formed relatively pure clusters: Cluster 5 contained 63.4% classical music, Cluster 1 contained 55.7% blues, and Cluster 4 contained 53.7% pop. These pure clusters demonstrate that features successfully capture distinct acoustic signatures for some genres. Conversely, other clusters mixed multiple genres: Cluster 2 combined country (18%), rock (17%), and blues (16%), reflecting the acoustic overlap between these related genres.

PCA visualization (Figure 4) reduced the 58-dimensional feature space to 2 dimensions, retaining 40.2% of total variance (PC1: 23.6%, PC2: 16.7%). The left panel shows true genre labels, revealing that metal and classical occupy distinct peripheral regions while blues, country, and jazz cluster together centrally. The right panel displays k-means cluster assignments with cluster centers marked as red crosses. The similarity between panels confirms that features naturally capture genre boundaries even without supervised training.

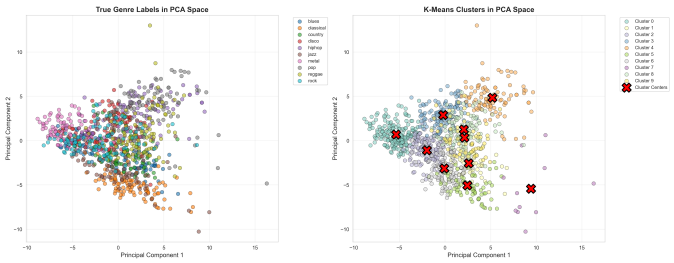

Fig. 4. PCA visualization of genre distribution in 2D feature space. Left: true genre labels showing natural separation. Right: k-means cluster assignments with centers marked as red X symbols. Metal and classical form distinct clusters while blues, jazz, and country overlap centrally, explaining classification challenges for these genres.

The cluster purity heatmap (Figure 5) quantifies the composition of each cluster, showing what percentage of each cluster belongs to each genre. Dark red cells indicate high purity (cluster dominated by single genre), while light yellow indicates mixture. Classical, metal, and pop show strong cluster purity, forming cohesive acoustic groups. In contrast, blues, jazz, country, and rock form mixed clusters, validating why these genres achieve lower classification accuracy.

This clustering analysis validates our supervised learning results: genres that cluster well unsupervised (classical, metal, pop) also achieve high classification accuracy in supervised learning, while genres that form mixed clusters (rock, disco, country) show lower accuracy. The consistency between supervised and unsupervised results provides strong evidence that our features effectively capture genre-discriminative acoustic properties, and that classification errors reflect genuine ambiguity in genre boundaries rather than feature deficiencies.

IV. Conclusions

This project successfully developed an automated music genre classification system achieving 77.5% accuracy on

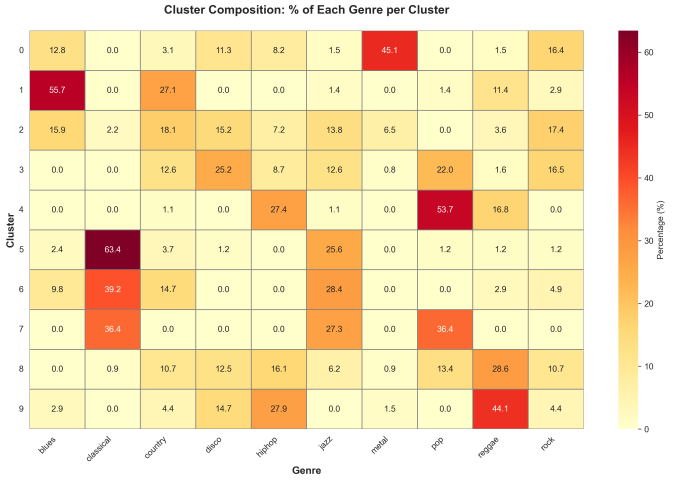

Fig. 5. Cluster purity heatmap showing the percentage composition of each k-means cluster. Each row represents one cluster, each column represents one genre. Dark red indicates high concentration (pure clusters), light yellow indicates mixing. Classical (Cluster 5: 63%), blues (Cluster 1: 56%), and pop (Cluster 4: 54%) form relatively pure clusters, while other genres mix significantly.

the GTZAN dataset, comparable to state-of-the-art results in published literature. Random Forest outperformed SVM (76.5%) and KNN (70.0%), demonstrating the effectiveness of ensemble methods for high-dimensional audio data with complex, non-linear decision boundaries.

Key findings include: (1) Classical, pop, and metal are most acoustically distinctive, achieving ¿85% accuracy due to unique signatures (orchestral arrangements, modern production, distorted guitars respectively). (2) Rock and disco are most challenging, with rock's internal diversity and disco's similarity to hip-hop causing frequent misclassifications. (3) Feature importance analysis revealed that chroma and MFCC features are most critical for genre discrimination, making the model interpretable rather than a "black box." We can explain which acoustic properties drive classification decisions. (4) K-means clustering validated that features naturally capture genre structure, with similar genres (blues-jazz-country) clustering together and distinct genres (classical-metal) separating, independent of supervised labels.

The primary limitation is the modest dataset size of 1,000 songs. Larger datasets would better represent intra-genre diversity and potentially improve accuracy, particularly for heterogeneous genres like rock. Additionally, the 30-second clip format may miss structural elements that define genres, such as verse-chorus progressions, bridges, and overall song development. Future work could explore: (1) Deep learning approaches using convolutional neural networks on mel-spectrograms for automatic feature learning, (2) recurrent neural networks to model temporal dynamics and song structure, (3) larger datasets like the Million Song Dataset or Spotify collections, (4) fine-grained classification into subgenres (e.g., death metal vs. thrash metal), and (5) multi-label classification recognizing that some songs span multiple genres.

Overall, this project demonstrates that machine learning can effectively automate genre classification with performance comparable to published research. The combination of supervised learning for classification, unsupervised validation for feature quality assessment, and explainability analysis through feature importance provides a comprehensive understanding of both the problem and solution. The systematic error patterns (e.g., blues-jazz confusion) reflect genuine acoustic similarities recognized even by human listeners, suggesting our system captures musically meaningful genre boundaries.

REFERENCES

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293-302, 2002.

[2] B. McFee et al., "librosa: Audio and music signal analysis in python," in *Proc. of the 14th Python in Science Conference*, 2015, pp. 18-25.

[3] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 303-319, 2011.

[4] B. L. Sturm, "The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use," *arXiv preprint arXiv:1306.1461*, 2013.

[5] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.