Music Genre Classification Using SVM and Genre Polyphony Discovery in Music

Ngô Hoài Nam and Hoàng Phạm Gia Bảo

FPT University, D1, Long Thanh My, Thu Duc, Ho Chi Minh, Vietnam
hmanclubs11@gmail.com, hoangbao9112005@gmail.com

Abstract. Rapid expansion of digital music libraries requires robust classification mechanisms to facilitate efficient retrieval, organization, and recommendation. As the volume of audio data increases exponentially, traditional manual classification methods become impractical. This study explores the application of Support Vector Machines (SVM) for automatic classification of music genres, a fundamental task in music information retrieval. The classification system utilizes supervised learning techniques to categorize audio tracks into predefined genres with high precision.

To achieve this, we extract and analyze key audio features, including Mel frequency spectral coefficients (MFCCs), spectral centroid, zero crossing rate, and chromagram,..., which serve as critical representations of timbre, rhythm, and harmonic content. The classification model is trained on the GTZAN dataset, a widely used benchmark for music genre classification, leveraging feature selection techniques and kernel optimization to enhance model performance and generalization.

Despite the inherent challenges of genre classification, such as overlapping genre characteristics, ambiguous genre boundaries, and the subjective nature of genre definitions, our optimized SVM model demonstrates the competitive classification precision. In addition, we discuss the limitations of SVM compared to deep learning-based approaches, highlighting potential areas for future improvements.

The SVM model achieves high accuracy in music genre classification; however, certain genres such as Rock and Hip-Hop face challenges due to overlapping audio characteristics. This reflects the inherent diversity and complexity of music, where genre boundaries are not strictly defined but often intersect, influenced by cultural, historical, and evolving listener preferences.

This research contributes to the advancement of machine learning applications in audio analysis and provides insights into the effectiveness of SVM for structured music categorization. The proposed approach has significant implications for various domains, including music recommendation systems, automated playlist generation, content-based retrieval, and intelligent audio tagging, ultimately improving user experience on digital music platforms and improving music information retrieval methodologies.

Keywords: Music Genre Classification, Support Vector Machine, Machine Learning, Audio Feature Extraction

1 Introduction

1.1 Background

The swift evolution of technology has sparked a change in the world of music altogether. With audio files and libraries increasing to unprecedented proportions, cataloging and retrieval has become a daunting task. Nowadays, users do not only require access to a sophisticated music library, they crave sophisticated services that help search and filter tracks based on their tastes. In such situations, automatic systems that categorize music into genres provide a powerful tool for classifying, tagging, and fast region browsing and retrieval of the music material.

1.2 Problem and Application

Problem: The problem of automatic music genre classification requires building a system capable of identifying the genre of a music segment or song based on its audio data. The system takes a digital music file (.wav) as input and outputs the corresponding music genre (Pop, Rock, Jazz, Classical, Hip Hop, etc.). The classification process typically involves key steps such as feature extraction (MFCC, Zero-Crossing Rate, Spectral Centroid, etc.) are included using the Librosa library, feature selection, training a machine learning model (SVM), model evaluation, and performance optimization.

Application: Automatic music genre classification has significant applications in various domains. In streaming services like Spotify and Apple Music, this technology helps recommend songs, generate automatic playlists, and improve search capabilities based on genre. For individual users, it enables the smart tagging and organization of personal music libraries. Furthermore, this technology supports market research, trend analysis, and popularity assessment of different genres. In education, it helps to analyze and compare musical compositions for learning purposes. The applications extend to audio recognition, content-based retrieval, and artificial intelligence research, contributing to advances in signal processing and machine learning for audio analysis.

1.3 Challenges and Difficulties

Automatic music genre classification presents various challenges due to the complexity and diversity of music. One of the biggest challenges is the subjectivity and ambiguity of genre classification. Genre boundaries are often unclear, evolving over time and across cultures, and a single song may incorporate elements from multiple genres, making accurate labeling difficult. In addition, genre perception varies between individuals, leading to inconsistencies in the training data.

Music is inherently complex, covering rhythm, melody, harmony, timbre, and structure. Extracting and selecting suitable features to represent these key aspects is a challenging task. Moreover, machine learning systems require large, well-labeled training datasets, but collecting and annotating such data is time-consuming and labor-intensive.

Another challenge is the "album effect" and the differences in recordings. Songs within the same album often share similar audio characteristics due to production and recording techniques, causing models to learn album-specific features rather than true genre traits. Selecting appropriate algorithms and optimizing models is also a complex problem that requires a balance between accuracy and generalization to avoid overfitting.

From a computational perspective, music genre classification models require significant resources, especially for feature extraction and training complex models. In real-time applications, such as music recommendation systems, optimization is necessary to ensure fast and accurate processing. Finally, handling new or rare genres, as well as music from diverse cultural backgrounds, poses a major challenge in developing a highly generalized classification system.

1.4 Project Objectives and Methodology

The primary goal of this project is to develop an automatic music genre classification system with high accuracy, capable of identifying and labeling music segments in various genres. To achieve this, the project follows a multistage process, including feature extraction, machine learning model application, and model optimization.

- First, crucial audio signal features such as MFCC, Zero-Crossing Rate, Spectral Centroid, Spectral Roll-off,... are extracted by using the Librosa library.

- Next, a machine learning algorithm, specifically Support Vector Machines (SVM), is used to train the classification model.

To ensure effective model performance, standard datasets such as GTZAN will be used for training and evaluation. Techniques such as feature selection, parameter optimization, kernel methods, and data augmentation will also be implemented to enhance model efficiency and generalization.

1.5 Expected Results and Contributions

The project aims to achieve high accuracy in the classification of music genres using advanced methods such as SVM with optimized kernels. Combining multiple audio features and employing suitable feature selection techniques will enhance classification performance. The music genre classification system will

be evaluated using standard datasets.

In addition, this research contributes to the field of music information retrieval, aiding the development of music recommendation systems, automatic music library organization, and music preference research. The proposed methods can also be extended to other tasks such as instrument classification, vocal recognition, and emotion analysis in music.The SVM model achieves high accuracy in music genre classification; however, certain genres such as Rock and Hip-Hop face challenges due to overlapping audio characteristics. This reflects the inherent diversity and complexity of music, where genre boundaries are not strictly defined but often intersect, influenced by cultural, historical, and evolving listener preferences

2 Method

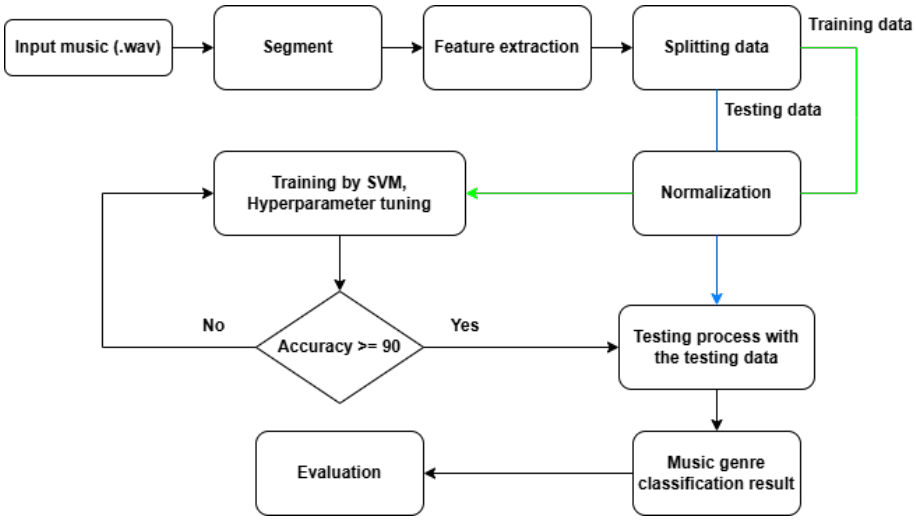

Fig. 1. Process flow diagram

2.1 Segment

Segmentation is an essential preprocessing step in music classification. By dividing an audio signal into ten equal parts, we effectively increase the number of training samples without requiring additional data. This enhances the model's ability to recognize patterns and improves classification performance.

Moreover, segmentation helps capture local features of the music, such as rhythm and harmony, which are crucial for distinguishing different genres. It

also reduces the complexity of each input sample, making the training process more efficient while minimizing the risk of overfitting.

Thus, segmentation not only increases data availability but also optimizes the learning process, contributing to more accurate and reliable music classification models.

2.2 Feature Extraction

The input to the Support Vector Machine (SVM) model is not the raw audio signal but a set of extracted features that encapsulate essential characteristics of music. These features serve as numerical representations that capturvarious timbral, harmonic, and rhythmic properties of the audio signal. In this study, we employ multiple feature representations, each contributing distinct insights into the underlying structure of the music signal. The selected features are as follows:

1. **Mel-Frequency Cepstral Coefficients (MFCCs)**: MFCCs are one of the most widely used features in audio signal processing, particularly in speech and music analysis [1]. They model the human auditory perception of sound frequencies by performing the following steps:
   (a) Applying the Short-Time Fourier Transform (STFT) to obtain the frequency domain representation of the signal [2].
   (b) Passing the magnitude spectrum through a set of Mel-scaled triangular filter banks to emphasize perceptually relevant frequency bands [3].
   (c) Computing the logarithm of the filter bank energies and applying the Discrete Cosine Transform (DCT) [4] to decorrelate the coefficients and obtain a compact representation.
   MFCCs effectively represent the spectral envelope of a signal and have been shown to be instrumental in various music classification tasks [5].

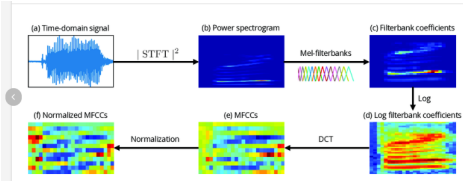

Fig. 2. Illustration of MFCC Feature Extraction Steps [6]

2. **Chromagram**: The chroma representation captures the distribution of energy across the 12 pitch classes (C, C#, D, ..., B), regardless of the octave. This feature provides insight into the harmonic content of music and is particularly useful for distinguishing genres with distinct harmonic structures [7].

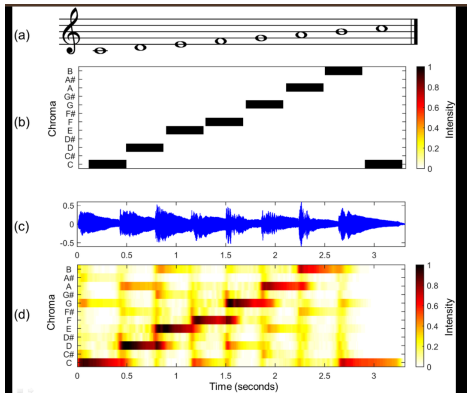

Fig. 3. Chroma Feature Representation of a Music Signal [8]

3. **Spectral Centroid**: The spectral centroid quantifies the "center of mass" of the frequency spectrum, serving as an indicator of the perceived brightness of a sound. It was first introduced by Peeters (2004) [9] and is mathematically

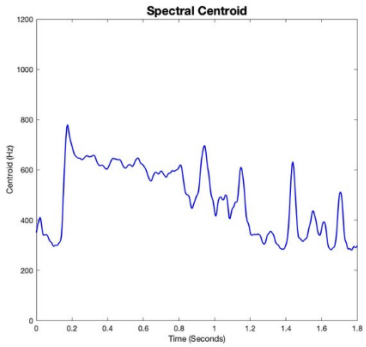

Fig. 4. Spectral Centroid Representation of an Audio Signal [10]

defined as:

$$SC = \frac{\sum_f f \cdot A(f)}{\sum_f A(f)} \tag{1}$$

where:

- f represents the frequency.

- $A(f)$ denotes the amplitude at frequency f.

A higher spectral centroid value indicates a brighter sound, typically found in instruments such as violins and trumpets, whereas lower values correspond to darker, bass-heavy sounds such as double bass and tubas [11].

4. **Spectral Rolloff**: Spectral rolloff is a measure of the frequency below which a certain percentage (typically 85%) of the total spectral energy is concentrated. It was introduced by Scheirer and Slaney (1997) [12]

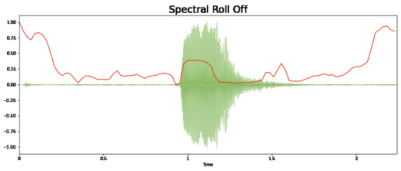

Fig. 5. Spectral Rolloff as an Indicator of Frequency Concentration [13]

defined as:

$$SR = \min_{f} \left(\sum_{i=0}^{f} A(i) \geq 0.85 \sum_{i=0}^{N} A(i) \right), \tag{2}$$

where:

- SR is the spectral roll-off point.
- f is the frequency bin index.
- $A(i)$ denotes the amplitude at frequency bin i.
- N represents the total number of frequency bins.
- 0.85 is a threshold that determines the roll-off percentage.

This feature helps in distinguishing between percussive and harmonic sounds. Higher spectral rolloff values indicate sounds with more high-frequency energy, such as cymbals and hi-hats, whereas lower values are characteristic of instruments like cellos and bass guitars.

5. **Zero-Crossing Rate (ZCR)**: ZCR quantifies the number of times a signal crosses the zero amplitude axis within a given time frame. It was introduced by Rabiner and Gold (1975) [14] and
is computed as:

$$ZCR = \frac{1}{N-1} \sum_{i=1}^{N-1} \mathbf{1}[x_i \cdot x_{i-1} < 0], \tag{3}$$

where:

- ZCR is the zero-crossing rate.
- N is the total number of samples.

- x_i represents the signal amplitude at time index i.
- $\mathbf{1}[\cdot]$ is the indicator function, which returns 1 if the condition inside is true and 0 otherwise.
- The condition $x_i \cdot x_{i-1} < 0$ checks if a zero-crossing occurs between two consecutive samples.

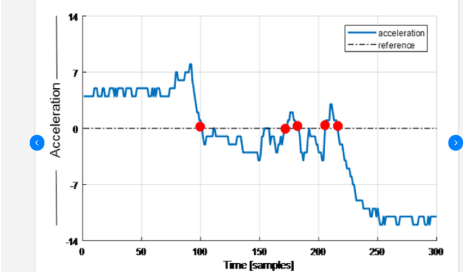

Fig. 6. Zero-Crossing Rate: A Measure of Signal Activity [15]

A higher ZCR indicates more rapid amplitude fluctuations, which are characteristic of percussive sounds such as snare drums and hi-hats, while lower ZCR values are typical for sustained harmonic sounds like violins or organ music.

6. **Tempo**: Tempo, measured in beats per minute (BPM), defines the speed of a musical piece. It is derived using onset detection and beat tracking algorithms, as proposed by Dixon (2001) [16]. Tempo plays a crucial role in genre classification, distinguishing fast-paced genres such as electronic dance music (EDM) and rock from slower genres such as jazz and ballads.

7. **Root Mean Square (RMS) Energy**
RMS energy measures the magnitude of the audio signal, providing insight into its loudness. It is defined as:

$$E_{RMS} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x_n^2} \tag{4}$$

- x_n represents the amplitude of the signal at sample n.
- N is the total number of samples.

RMS energy has been widely studied in audio signal processing.

8. **Harmony**
Harmony represents the consonance and dissonance in an audio signal, often derived from harmonic-to-noise ratio (HNR). It is computed as:

$$H = 10 \log_{10}\left(\frac{P_H}{P_N}\right) \tag{5}$$

- P_H is the power of the harmonic components.
- P_N is the power of the noise components.

Harmony analysis has been extensively studied in music perception research [17].

9. **Perceptual Spectral Bandwidth**
Spectral bandwidth quantifies the spread of spectral energy and is crucial for timbre analysis. It is defined as:

$$BW_p = \left(\sum_{k=1}^{K} |f_k - f_c|^p S_k \right)^{\frac{1}{p}} \tag{6}$$

- f_k is the frequency of bin k.
- f_c is the spectral centroid.
- S_k is the spectral magnitude at bin k.
- p is typically set to 2.

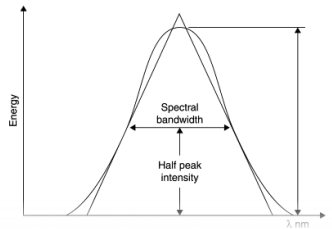

Fig. 7. Visualization of spectral bandwidth, illustrating the distribution of spectral energy around the spectral centroid. This feature plays a crucial role in timbre analysis and music classification.

[18]

This feature has been analyzed in auditory perception studies[9].

10. **Percussion**
Percussion features capture the rhythmic and transient characteristics of a musical signal. One common approach to measure percussiveness is through onset strength, defined as:

$$O(n) = \sum_{k=1}^{K} H(S_k(n) - S_k(n-1)) \tag{7}$$

- $S_k(n)$ is the spectral magnitude at bin k and time frame n.
- $H(.)$ is the Heaviside step function.

High onset strength indicates strong percussive events in a signal [19].

Each audio file is transformed into a feature vector where each dimension corresponds to a specific extracted feature. For example, when 13 MFCC coefficients are used alongside chroma, spectral, and rhythmic features, the final feature vector may comprise 18 to 20 dimensions, depending on the selected representation. These extracted features serve as input to the SVM classifier, enabling it to learn meaningful patterns for music genre classification.

2.3 Support Vector Machine (SVM) Classification

Training Phase: SVM is a supervised learning algorithm that constructs an optimal hyperplane for separating different music genres. Given a training dataset $\{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^d$ represents the extracted feature vector and $y_i \in \{-1, 1\}$ denotes the genre label, SVM solves the following optimization problem as introduced by Vapnik (1995) [20]:

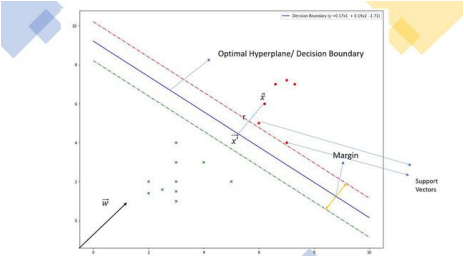

Fig. 8. SVM Decision Boundary with Support Vectors [21]

$$\min_{w,b} \frac{1}{2}\|w\|^2 \tag{8}$$

$$\text{subject to} \quad y_i(w^T x_i + b) \geq 1, \quad \forall i \tag{9}$$

Where:

- w is the weight vector, which determines the orientation of the decision boundary.
- b is the bias term, which shifts the decision boundary without altering its orientation.
- y_i is the class label of the data point x_i, taking values in $\{-1, 1\}$.
- x_i is the feature vector representing a data point in the input space.

The model aims to maximize the margin between classes while minimizing classification error. The hyperplane is selected such that it maximizes the distance from the nearest data points in each class, ensuring better generalization on unseen data.

For non-linearly separable data, the kernel trick is employed to project data into a higher-dimensional space using transformation $\phi(x)$. This allows the SVM to classify data that is not linearly separable in the original space [22].

Classification Phase: Once trained, the SVM model classifies new music samples by computing:

$$y = \text{sign}(w^T x + b) \tag{10}$$

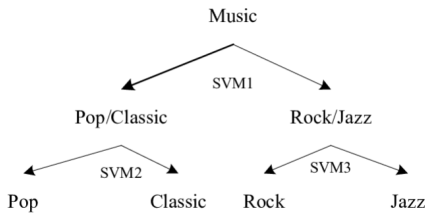

Fig. 9. Hierarchical Support Vector Machine Approach for Music Genre Classification [23]

where:
• y is the predicted class label.
• x is the feature vector of the test sample.
• w is the weight vector.
• b is the bias term.
• sign$(\cdot)$ is the sign function, which returns $+1$ if the input is positive and -1 if the input is negative.

If probability estimates are required, Platt Scaling [24] can be applied to convert SVM outputs into probabilistic confidence scores.

Radial Basis Function (RBF) Kernel: The **Radial Basis Function (RBF) Kernel** is one of the most widely used kernels in **Machine Learning** models, especially in **Support Vector Machine (SVM)**. The general formula of the RBF kernel is defined as follows:

$$k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \tag{11}$$

where:
• x_i, x_j are two input vectors in the feature space.
• $\|x_i - x_j\|^2$ is the squared Euclidean distance between the two vectors.
• γ (gamma) is a parameter that controls the influence of the distance between data points.
• exp$(\cdot)$ is the exponential function with base **e**, ensuring that the kernel value

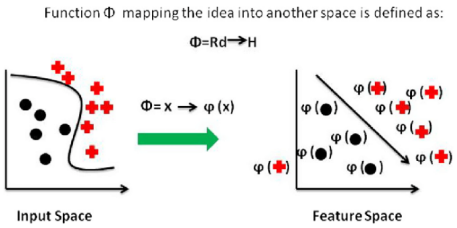

Fig. 10. Mapping from Input Space to Feature Space using Radial Basis Function (RBF) Kernel

[25]

remains within the range $(0, 1]$.

The RBF kernel transforms data from the original space into a higher-dimensional space, where the samples can become linearly separable. Specifically:
• When γ is large, the distance between two points x_i and x_j has a strong influence, making the model more complex and prone to overfitting.
• When γ is small, the distance between two points has less impact, making the model more generalized but potentially less flexible in identifying complex decision boundaries.

The RBF kernel is widely used in classification and regression tasks such as:
• Non-linear data classification using **Support Vector Machine (SVM)**.
• Pattern recognition and image processing.
• Unsupervised learning in **Gaussian Processes**.

Due to its flexibility and efficiency, the RBF kernel is a powerful choice for many real-world machine learning problems.

3 Related works

The classification of music genres using machine learning has been an active area of research, with Support Vector Machines (SVM) emerging as a prominent technique due to its robustness in high-dimensional feature spaces. Several studies have explored the effectiveness of SVM for music classification, demonstrating its advantages and limitations in various contexts.

3.1 Advantages and Limitations of SVM in Music Classification

SVM has been widely recognized for its strong generalization capabilities, particularly in high-dimensional spaces. Its ability to implicitly operate in a transformed feature space allows it to capture complex nonlinear relationships within music data. Key advantages of SVM include:
Advantages of SVM:

• **Robust generalization**: SVM optimizes a bound on the generalization error, mitigating the curse of dimensionality [26].
• **Kernel-based flexibility**: By employing kernel functions, SVM effectively captures nonlinear structures in music data [27].
• **Computational efficiency**: Compared to Gaussian Mixture Models (GMM) and K-Nearest Neighbors (KNN), SVM offers efficient classification, particularly with small test sets [11].
• **Scalability**: Unlike KNN, which relies on the entire training dataset, SVM depends only on support vectors, leading to reduced computational complexity [28].
• **Feature selection capability**: SVM implicitly assesses the relevance of extracted features, facilitating more effective feature engineering [29].
• **Balanced complexity control**: SVM enables a trade-off between model complexity and generalization, making it suitable for real-time applications such as beat tracking and automatic accompaniment systems [11]. Nevertheless, SVM also exhibits several limitations:
Disadvantages of SVM:
• **High computational cost**: Training an SVM model can be resource-intensive, particularly for large datasets [30].
• **Hyperparameter sensitivity**: The performance of SVM is highly dependent on kernel selection and parameter tuning [31].
• **Scalability challenges**: While effective for small to medium-sized datasets, SVM may struggle with scalability in big data applications [24].

3.2 Impact of Kernel Functions on SVM Performance

The incorporation of kernel functions enhances SVM's flexibility and generalization ability. Notable benefits of kernel methods include:
Advantages of Kernelized SVM:
• **Increased adaptability**: Kernel functions allow SVM to operate effectively in high-dimensional spaces, capturing complex feature relationships [27].
• **Enhanced classification performance**: Kernelized SVMs optimize decision boundaries, reducing the risk of overfitting in intricate datasets [30].
• **Superior accuracy**: Empirical studies have demonstrated that kernel-based SVMs outperform traditional linear classifiers in non-linearly separable data settings [28].
• **Wide applicability**: Kernelized SVMs have found applications in diverse fields, including text classification, bioinformatics, speaker recognition, and various MIR tasks [26].

3.3 Review of Prior Studies in Music Genre Classification

Several studies have explored different machine learning techniques for music genre classification, with SVM serving as a fundamental approach:

Related Works:
Several studies have explored different machine learning techniques for music genre classification, with SVM serving as a fundamental approach.

Tzanetakis and Cook (2002) introduced the widely used GTZAN dataset and applied machine learning techniques for genre classification [11]. Later, Das and Kolya (2019) implemented a single-layer feedforward neural network for genre classification, achieving an accuracy of 84.3% [32]. In a similar effort, Dong (2018) employed convolutional neural networks (CNN) and achieved human-level classification accuracy [33].

Zhang et al. (2016) proposed an improved CNN-based model that achieved 84. 8% and 87. 4% precision, demonstrating enhanced performance compared to previous deep learning approaches [34]. Meanwhile, Chen et al. (2010) explored Dynamic Frame Analysis with SVM, reaching an impressive accuracy of 98% on a subset of six genres [35]. More recently, Chaudary et al. (2021) introduced Empirical Mode Decomposition (EMD) preprocessing combined with SVM, achieving precision of 94. 0% across five music genres [36]. Furthermore, Patil et al. (2017) evaluated multiple feature sets with SVM and found that the polynomial kernel SVM performed best, achieving an accuracy of 78% [37].

These studies highlight the competitiveness of SVM in the classification of music genres, particularly when integrated with appropriate feature extraction and kernel selection techniques. Although deep learning methods have gained prominence in recent years, SVM remains a viable alternative, especially in scenarios with computational and data constraints.

3.4 Trends in Music Genre Classification Research

The evolution of music genre classification research can be categorized into distinct phases: **Music Genre Classification Trends:**

• **Pre-2020**: Dominated by traditional machine learning techniques such as SVM, KNN, and GMM, utilizing hand-crafted features such as Mel frequency cepstral coefficients (MFCC) and spectral characteristics.
• **2020-2023**: Marked by a paradigm shift towards deep learning models, including CNNs, recurring neural networks (RNNs), and hybrid approaches, alongside advancements in preprocessing methods such as EMD.
• **2024 and Beyond**: Characterized by the integration of Transformer-based architectures, ensemble learning strategies, and multimodal data fusion to enhance classification accuracy and interpretability.

In summary, while the field of music genre classification has evolved significantly, SVM, particularly with kernel enhancements, remains a crucial tool. The growing adoption of deep learning has yielded notable improvements in classification performance; however, SVM continues to offer competitive advantages in computationally constrained settings and applications requiring interpretable decision boundaries.

4 Experiments

4.1 Dataset Description

The dataset used in this study is the **GTZAN** dataset, a standard benchmark for music genre classification, introduced by George Tzanetakis and Perry Cook in 2002 [11]. This dataset consists of 1,000 audio clips, each lasting 30 seconds, evenly distributed across 10 different genres. The table below provides the genre labels and their corresponding encoded values:

Genre	Encoded Label
Blues	0
Classical	1
Country	2
Disco	3
Hip-Hop	4
Jazz	5
Metal	6
Pop	7
Reggae	8
Rock	9

Table 1. Genre labels and corresponding encoded values.

Each genre contains 100 audio samples, ensuring a balanced dataset. To enhance the number of training samples, each 30-second clip is segmented into 10 non-overlapping **3-second** segments, resulting in a total of **10,000** data samples. This approach enables the model to capture finer musical characteristics while mitigating the impact of long-term signal variations.

The key extracted audio features include:

- **Chroma Features**: Represent pitch content.
- **RMS Energy**: Measures signal intensity.
- **Spectral Features** (Centroid, Bandwidth, Rolloff): Describe the spectral characteristics of the signal.
- **Zero-Crossing Rate (ZCR)**: Counts the number of times the signal crosses zero amplitude.
- **Mel-Frequency Cepstral Coefficients (MFCCs)**: Extracts perceptual frequency-domain features.
- **Harmony and Percussive Components**:
  - **Harmony**: Represents the harmonic content of the signal, capturing melody and chord structure.
  - **Percussive**: Captures percussive elements, such as drums and rhythmic structures.
- **Tempo**: Estimates the beats per minute (BPM) of the track.

4.2 Experimental Setup

Data Preprocessing Prior to training, all extracted features are standardized using **StandardScaler** to ensure a mean of 0 and a standard deviation of 1:

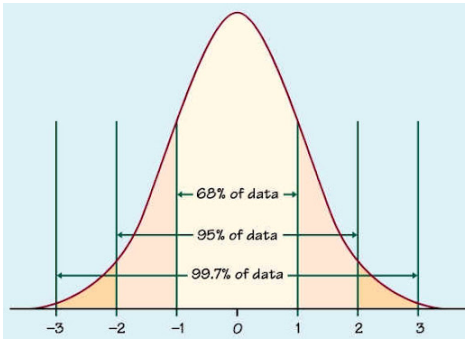

Fig. 11. Z-score histogram[38]

$$Z_{\text{scaled}} = \frac{X - \mu}{\sigma} \tag{12}$$

- μ is the mean of the dataset.
- σ is the standard deviation of the dataset.

Data Splitting The dataset is split into **80% training** and **20% testing** using the **train-test split** method, ensuring no overlap between the training and test sets.

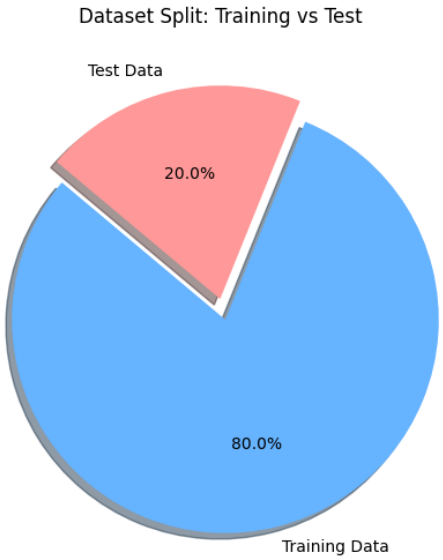

Fig. 12. Data distribution

Model Training The classifier used in this experiment is the **Support Vector Machine (SVM)**, known for its effectiveness in high-dimensional spaces. To optimize performance, **GridSearchCV** is applied to find the best hyperparameters.

GridSearchCV systematically explores a predefined hyperparameter space using cross-validation. The general formula for **cross-validation** in GridSearchCV is given by:

$$\text{Score}_{cv} = \frac{1}{k} \sum_{i=1}^{k} \text{Score}_i \tag{13}$$

where:

- k is the number of folds in cross-validation.
- Score_i represents the performance score on the i-th validation set.

This method ensures robust model evaluation and prevents overfitting.

4.3 Experimental Results

The performance of the trained model on the test set is summarized as follows:

- **Overall Accuracy**: **90.9%**
- **Macro-averaged Precision**: **91%**
- **Macro-averaged Recall**: **91%**
- **Macro-averaged F1-score**: **91%**

The class-wise classification report is shown in Table 2.

Encoded label	Genre	Precision	Recall	F1-score
0	Blues	0.88	0.91	0.90
1	Classical	0.98	0.95	0.96
2	Country	0.88	0.86	0.87
3	Disco	0.97	0.90	0.93
4	Hip-Hop	0.78	0.93	0.85
5	Jazz	0.92	0.99	0.96
6	Metal	0.97	0.95	0.96
7	Pop	0.90	0.94	0.92
8	Reggae	0.92	0.93	0.93
9	Rock	0.91	0.80	0.85

Table 2. Class-wise precision, recall, and F1-score.

From the table, it can be observed that genres such as **Classical, Jazz, and Metal** achieve high classification performance, while **Hip-Hop and Rock** exhibit slightly lower scores due to their complex spectral properties.

Below are visualizations of the classification results:

Fig. 13. Confusion matrix

4.4 Experimental Analysis on Specific Samples

1. **Samples from the Test Set** These samples are selected directly from the test data set to evaluate the performance of the model under predefined conditions.

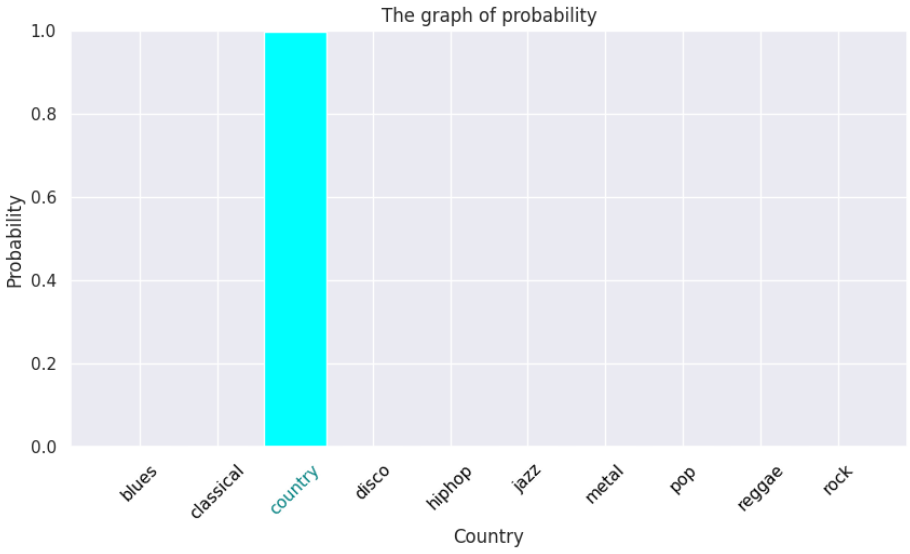

Fig. 14. Good performance

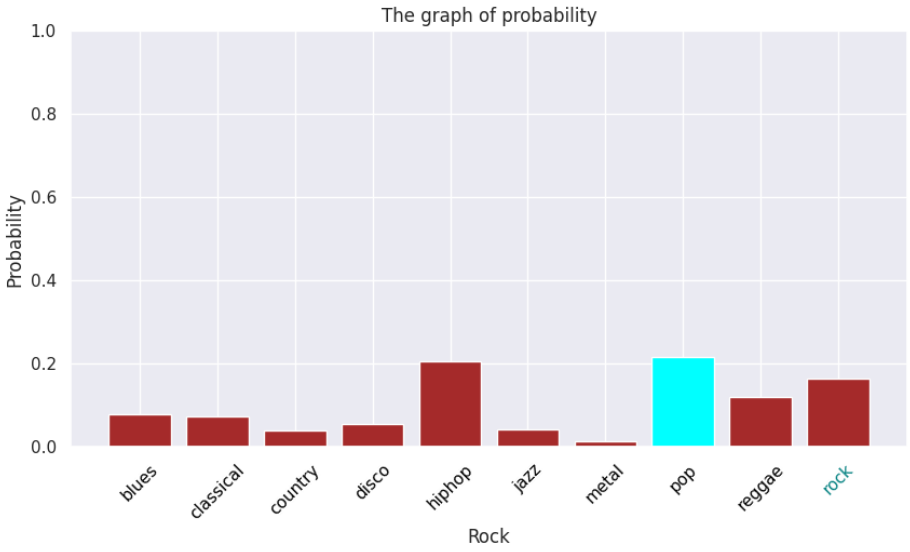

Fig. 15. Bad performance

2. **Samples with a Specific Genre**
   This category consists of samples belonging to a distinct and well-defined genre, allowing for an in-depth analysis of model accuracy in genre classification.

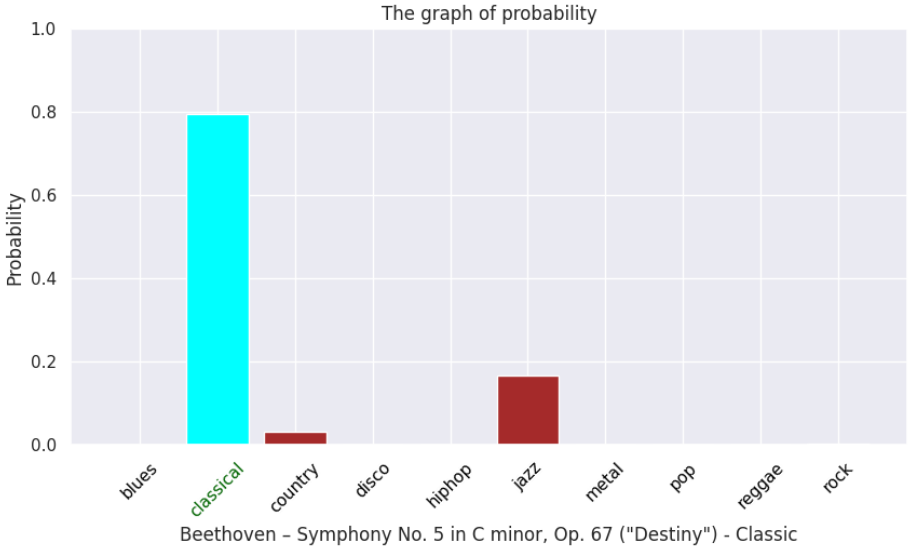

Fig. 16. Beethoven – Symphony No. 5 in C minor, Op. 67 ("Destiny") - Classic

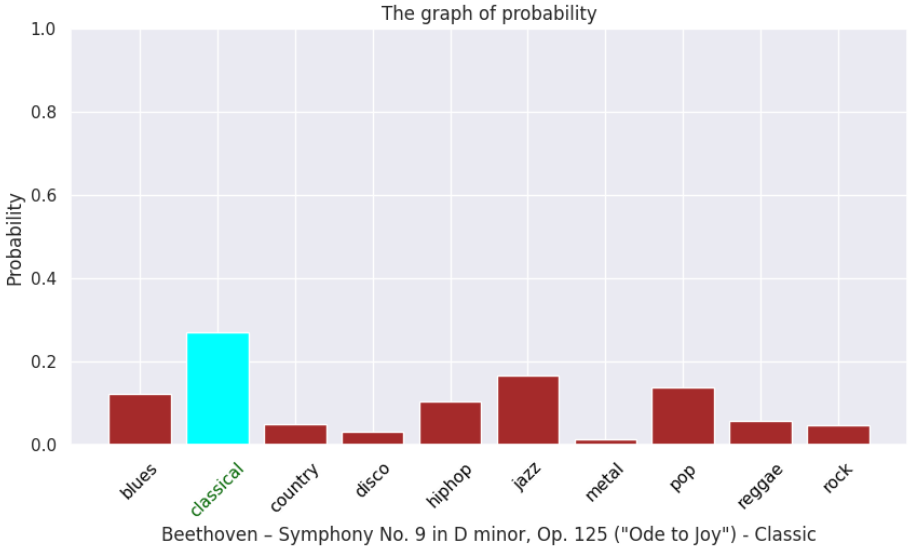

Fig. 17. Beethoven – Symphony No. 9 in D minor, Op. 125 ("Ode to Joy") - Classic

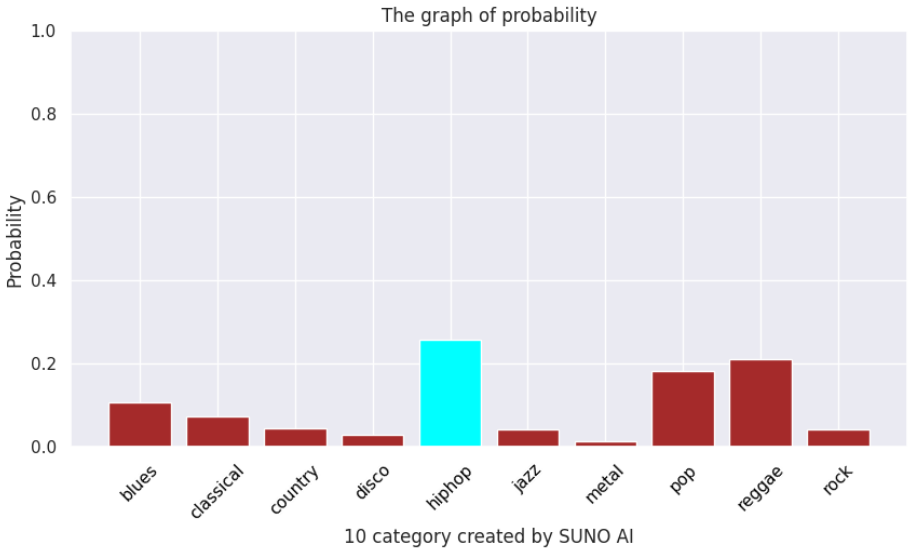

Fig. 18. 10 category mixed together

3. **The samples are mixed from many genres (Modern Music) and are untrained**
   These samples include components from various genres and are untrained, with a particular focus on modern music. Analyzing such samples allows to evaluate the model's ability to handle genre combinations and detect complex patterns in mixed compositions.

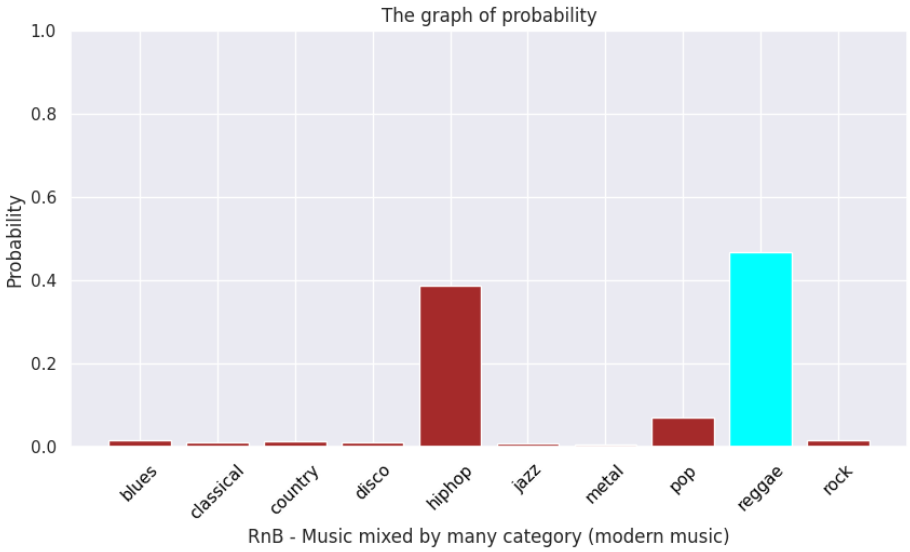

Fig. 19. Xích Thêm Chút - XTC Remix | RPT Groovie ft TLinh x RPT MCK (Prod. by fat_benn & RPT LT)| RAPITALOVE - RnB

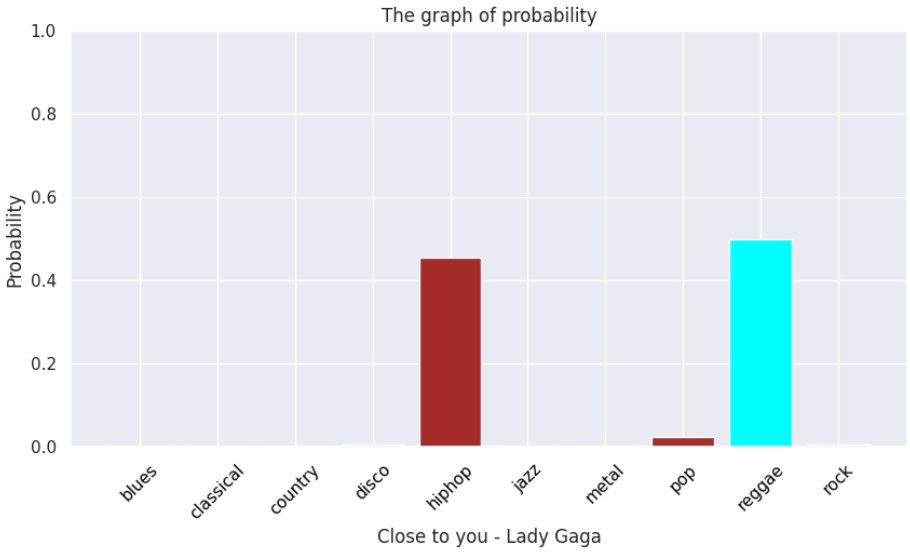

Fig. 20. Lady Gaga - Close To You - Unknowed category

4.5 Error Analysis and Investigation

Through empirical experimentation:

- **Performance on the Test Set Samples:**
  The model demonstrates accurate and reliable predictions across most genres, except for Hip-Hop and Rock, where its performance is not yet optimal.
- **Samples with a Specific Genre:**
  When evaluating specific genres, the model correctly identifies classical music through two renowned works by Beethoven: *Symphony No. 5 in C minor, Op. 67* ("Destiny") and *Symphony No. 9 in D minor, Op. 125* ("Ode to Joy"). However, for *Symphony No. 9 in D minor, Op. 125*, the model's performance is not as robust as for *Symphony No. 5 in C minor, Op. 67*, prompting further inquiry. This raises the question of whether *Symphony No. 9* is purely classical or contains elements of multiple genres.
  To investigate this hypothesis, I examined whether the model could reliably classify a piece containing multiple genres. As shown in Figure 18, I utilized SUNO AI—an AI-driven generative music system—to generate ten distinct musical genres that were part of the training dataset. The results revealed a relatively even distribution of these genres, as illustrated in Figure 18.
  To further validate this, I tested the model on a musical piece that had been mixed from multiple genres but had not been encountered during training. The results indicated that for a sample incorporating various genres (modern music), the model identified a blend of different musical styles rather than a single, fixed genre. Notably, Hip-Hop and Reggae were among the most extensively mixed genres within the composition.

5 Conclusion

The automatic classification of music genres remains a challenging task due to the complex and evolving nature of musical compositions. In this study, we explored the application of Support Vector Machines (SVM) for genre classification, leveraging key audio features such as Mel-Frequency Cepstral Coefficients (MFCCs), spectral features, and rhythmic descriptors. Through extensive experimentation, we demonstrated that SVM, particularly with kernel optimizations, provides a robust and computationally efficient approach to classifying music genres.

Our results indicate that segment-based data partitioning significantly improves classification accuracy by enhancing the model's ability to capture local musical characteristics. By increasing the number of training samples, segmentation mitigates overfitting and improves generalization across diverse musical styles. Additionally, hyperparameter tuning and feature selection played a critical role in optimizing model performance.

Despite achieving competitive accuracy, challenges remain in handling ambiguous genre boundaries, multi-genre compositions, and variations introduced by different recording techniques. Future research could focus on integrating deep learning architectures such as Convolutional Neural Networks (CNNs) with traditional SVM approaches to further enhance classification accuracy. Additionally, exploring attention-based mechanisms and hybrid feature representations may provide deeper insights into the temporal and harmonic structures of music.

In conclusion, this research contributes to the advancement of music information retrieval by demonstrating the effectiveness of SVM in genre classification. The findings provide a foundation for future developments in automated music tagging, recommendation systems, and intelligent audio processing applications.The SVM model achieves high accuracy in music genre classification; however, certain genres such as Rock and Hip-Hop face challenges due to overlapping audio characteristics. This reflects the inherent diversity and complexity of music, where genre boundaries are not strictly defined but often intersect, influenced by cultural, historical, and evolving listener preferences.

Acknowledgements

We would like to express our deepest gratitude to our advisor, Dr. Đỗ Đức Hào, for his invaluable guidance, insightful feedback, and continuous support throughout the research process. His expertise and encouragement have been instrumental in shaping this study.

We are also grateful to our peers and colleagues for their constructive discussions and support, as well as to our families and friends for their unwavering encouragement.

Finally, we extend our appreciation to all participants who contributed to the survey, providing essential data for this research.

Thank you all for your support and inspiration.

References

[1] Steven B. Davis and Paul Mermelstein. "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28.4 (1980), pp. 357–366.

[2] J. W. Cooley and J. W. Tukey. "An algorithm for the machine calculation of complex Fourier series". In: *Mathematics of Computation* 19.90 (1965), pp. 297–301.

[3] J. Volkmann S. S. Stevens and E. B. Newman. "A scale for the measurement of the psychological magnitude pitch". In: *The Journal of the Acoustical Society of America* (1937).

[4] T. Natarajan N. Ahmed and K. R. Rao. "Discrete Cosine Transform". In: *IEEE Transactions on Computers* 23.1 (1974), pp. 90–93.

[5] Beth Logan. "Mel Frequency Cepstral Coefficients for Music Modeling". In: *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*. 2000.

[6] ResearchGate. *The process of computing Mel-Frequency Cepstral Coefficients (MFCCs)*. Accessed: 2025-03-16. 2024. URL: https://www.researchgate.net/figure/The-process-of-computing-Mel-Frequency-Cepstral-Coefficients-MFCCs-observed-in-Figure_fig1_373016283.

[7] H. Fujishima. "Realtime Chord Recognition of Musical Sound: A System Using Common Lisp Music". In: *Proceedings of the International Computer Music Conference (ICMC)*. 1999.

[8] Academia.edu. *Chroma Feature Extraction*. Accessed: 2025-03-16. 2020. URL: https://www.academia.edu/42216949/Chroma_Feature_Extraction.

[9] G. Peeters. *A Large Set of Audio Features for Sound Description (Similarity and Classification) in the CUIDADO Project*. Tech. rep. IRCAM, 2004.

[10] Timbre and Orchestration. *Spectral Centroid*. Accessed: 2025-03-16. 2019. URL: https://timbreandorchestration.org/writings/timbre-lingo/2019/3/29/spectral-centroid.

[11] G. Tzanetakis and P. Cook. "Musical genre classification of audio signals". In: *IEEE Transactions on Speech and Audio Processing* 10.5 (2002), pp. 293–302.

[12] E. Scheirer and M. Slaney. "Construction and evaluation of a robust multifeature speech/music discriminator". In: (1997), pp. 1331–1334.

[13] ResearchGate. *Representation of Spectral Roll-off feature of Sound wave of Ko-5 Chroma Frequency*. Accessed: 2025-03-16. 2021. URL: https://www.researchgate.net/figure/Representation-of-Spectral-Roll-off-feature-of-Sound-wave-of-Ko-5-Chroma-Frequency-To_fig4_352357606.

[14] L. R. Rabiner and M. R. Sambur. "An algorithm for determining the endpoints of isolated utterances". In: *The Bell System Technical Journal* 54.2 (1975), pp. 297–315.

[15] ResearchGate. *Zero-crossing rate (ZCR) feature depicted for x-component of brush teeth signal*. Accessed: 2025-03-16. 2020. URL: https://www.researchgate.net/figure/Zero-crossing-rate-zcr-feature-depicted-for-x-component-of-brush-teeth-signal_fig4_337921090.

[16] Simon Dixon. "Automatic extraction of tempo and beat from expressive performances". In: *Journal of New Music Research* 30.1 (2001), pp. 39–58.

[17] Paul Boersma. "Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound". In: *Proceedings of the Institute of Phonetic Sciences* 17 (1993), pp. 97–110. URL: https://fon.hum.uva.nl/paul/papers/Proceedings_1993.pdf.

[18] ScienceDirect. *Spectral Bandwidth*. Accessed: 2025-03-16. 2025. URL: https://www.sciencedirect.com/topics/engineering/spectral-bandwidth.

[19] Juan Pablo Bello et al. "A Tutorial on Onset Detection in Music Signals". In: *IEEE Transactions on Speech and Audio Processing* 13.5 (2005),

pp. 1035–1047. DOI: 10.1109/TSA.2005.851998. URL: https://www.iro.umontreal.ca/~pift6080/H09/documents/papers/Bello-TSAP-2005.pdf.

[20] C. Cortes and V. Vapnik. "Support-vector networks". In: *Machine Learning* 20.3 (1995), pp. 273–297.

[21] GeeksforGeeks. *Separating Hyperplanes in SVM*. Accessed: 2025-03-16. 2025. URL: https://www.geeksforgeeks.org/separating-hyperplanes-in-svm/.

[22] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. "Kernel Principal Component Analysis". In: *Artificial Neural Networks – ICANN 1997*. Springer, 1997, pp. 583–588.

[23] ResearchGate. *Musical genre classification diagram - Support Vector Machine (SVM)*. Accessed: 2025-03-16. 2005. URL: https://www.researchgate.net/figure/Musical-genre-classification-diagram-41-Support-Vector-Machine-Support-vector-machine_fig2_4015150.

[24] John C. Platt. "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods". In: *Advances in Large Margin Classifiers* (1999), pp. 61–74.

[25] ResearchGate. *Representation of Radial Basis Function (RBF) kernel Support Vector Machine*. Accessed: 2025-03-16. 2019. URL: https://www.researchgate.net/figure/Representation-of-Radial-basis-function-RBF-kernel-Support-Vector-Machine_fig2_329807137.

[26] Vladimir N. Vapnik. *Statistical Learning Theory*. New York: Wiley, 1998.

[27] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press, 2002.

[28] Christopher J.C. Burges. "A Tutorial on Support Vector Machines for Pattern Recognition". In: *Data Mining and Knowledge Discovery* 2.2 (1998), pp. 121–167. DOI: 10.1023/A:1009715923555.

[29] Isabelle Guyon et al. "Gene Selection for Cancer Classification using Support Vector Machines". In: *Machine Learning* 46.1-3 (2002), pp. 389–422. DOI: 10.1023/A:1012487302797.

[30] S. S. Keerthi et al. "Improvements to Platt's SMO Algorithm for SVM Classifier Design". In: *Neural Computation* 13.3 (2001), pp. 637–649. DOI: 10.1162/089976601300014493.

[31] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. *A Practical Guide to Support Vector Classification*. Tech. rep. Department of Computer Science, National Taiwan University, 2003. URL: https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

[32] Sourav Das and Anup Kumar Kolya. "Detecting Generic Music Features with Single Layer Feedforward Network using Unsupervised Hebbian Computation". In: *arXiv preprint arXiv:2008.13609* (2019). URL: https://arxiv.org/pdf/2008.13609.pdf.

[33] Mingwen Dong. "Convolutional Neural Network Achieves Human-level Accuracy in Music Genre Classification". In: *arXiv preprint arXiv:1802.09697* (2018). URL: https://arxiv.org/abs/1802.09697.

[34] Liang Zhang et al. "Improved Music Genre Classification with Convolutional Neural Networks". In: *Proceedings of Interspeech 2016*. 2016, pp. 3304–3308. URL: https://www.isca-speech.org/archive/Interspeech_2016/pdfs/1360.PDF.

[35] Liang Chen, Li Su, and Yi-Hsuan Yang. "Music Genre Classification Algorithm Based on Dynamic Frame Analysis and Support Vector Machine". In: *Proceedings of the 2010 IEEE International Symposium on Multimedia (ISM)*. 2010, pp. 357–362. DOI: 10.1109/ISM.2010.65.

[36] Eamin Chaudary, Paul Gretschmann, and Sumair Aziz. "Music Genre Classification using Support Vector Machine and Empirical Mode Decomposition". In: *2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC)*. 2021, pp. 357–362. DOI: 10.1109/MAJICC53071.2021.9526251. URL: https://www.researchgate.net/profile/Sumair-Aziz/publication/354411895_Music_Genre_Classification_using_Support_Vector_Machine_and_Empirical_Mode_Decomposition/links/63f32a2731cb6a6d1d174073/Music-Genre-Classification-using-Support-Vector-Machine-and-Empirical-Mode-Decomposition.pdf.

[37] Nilesh M. Patil and Milind U. Nemade. "Music Genre Classification Using MFCC, K-NN and SVM Classifier". In: *International Journal of Computer Engineering In Research Trends* 4.2 (2017), pp. 43–47. ISSN: 2349-7084. URL: https://www.ijcert.org/.

[38] Larry Green. *Z-Score*. Accessed: 2025-03-16. 2025. URL: http://www.ltcconline.net/greenl/courses/201/probdist/zScore.htm.

This research was carried out using the GTZAN dataset, which includes 1,000 music clips spanning ten genres.[1]

[1] GTZAN dataset: A widely recognized dataset for music genre classification, first introduced by Tzanetakis and Cook (2002). Each genre contains 100 tracks, used extensively in audio signal processing and machine learning research.