UNIVERSITÀ
DEGLI STUDI
DI MILANO

DIPARTIMENTO DI
INFORMATICA

Audio Pattern Recognition

Comprehensive Comparison of CNN, RNN, SVM, and Random Forest Classifiers for Music Genre Recognition Using MFCCs

Author: Moein Ghaeini H.

Matriculation Number: 14460A

Program: Master's Degree in Computer Science

Prof. Stavros Ntalampiras

Academic Year: 2024/2025

June 9, 2025

Abstract

This study explores and compares the effectiveness of machine learning models in music genre classification using the GTZAN Genre Collection dataset. The models considered include Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Support Vector Machines (SVM), and Random Forests (RF), with Mel-frequency cepstral coefficients (MFCCs) as the key audio feature. The study also evaluates the effect of K-means clustering as a preprocessing step and provides insights through confusion matrices, accuracy curves, and classification reports.

1 Introduction

Music genre classification represents a fundamental task within the field of audio pattern recognition, with important applications in music recommendation systems, automated tagging, and the organization of digital music libraries. The task is inherently complex due to the temporal variability and high dimensionality of audio signals. Traditional classification approaches have relied on manually engineered features—such as pitch, rhythm, and timbre—combined with conventional machine learning algorithms like Support Vector Machines (SVMs) and decision trees. While these methods required substantial domain expertise, they often fell short in capturing the intricate patterns present in musical data.Recent advancements in deep learning have significantly improved genre classification performance by enabling end-to-end learning from raw or minimally processed features. Convolutional Neural Networks (CNNs) have proven effective in extracting spatial patterns from time-frequency representations, whereas Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) architectures, are adept at modeling temporal dependencies in sequential data.This study conducts a comparative analysis of CNN and RNN models using Mel-frequency cepstral coefficients (MFCCs) as input features. The models are evaluated based on classification accuracy, training efficiency, and architectural complexity. Additionally, the study explores K-means clustering as an unsupervised alternative to investigate the potential for feature-based genre grouping. To provide a comprehensive performance benchmark, traditional classifiers such as SVMs and Random Forests (RFs) are also assessed using the same MFCC-based features derived from the GTZAN dataset.

2 Dataset and Feature Extraction

2.0.1 Dataset and Feature Extraction

GTZAN Genre Collection The GTZAN Genre Collection is a widely used benchmark dataset in music information retrieval. It comprises 1,000 audio tracks, evenly distributed across 10 distinct genres: Blues, Classical, Country, Disco, Hip-hop, Jazz, Metal, Pop, Reggae, and Rock. Each track has a duration of 30 seconds and is sampled at 22,050 Hz. To enhance data volume and increase diversity, each track was segmented into non-overlapping 3-second clips, resulting in a total of 10,000 audio samples.

Audio Preprocessing and Segmentation Prior to feature extraction, all audio tracks were normalized to a consistent amplitude range. Each 30-second track was then divided into non-overlapping 3-second segments. This segmentation strategy not only expanded the dataset size but also introduced greater variability in training data, thereby improving model generalization.

MFCC Feature Extraction Mel-Frequency Cepstral Coefficients (MFCCs) were extracted from each 3-second audio segment using the Librosa library. Thirteen MFCCs were computed per segment through the following procedure:

- **Signal Resampling**: All signals were standardized to a sampling rate of 22,050 Hz to ensure uniformity across samples.
- **Framing and Windowing**: Each segment was divided into overlapping frames of 2048 samples with a hop length of 512. A Hamming window was applied to each frame to reduce spectral leakage. These operations were managed internally via the `librosa.feature.mfcc` function.
- **Spectral Transformation**: A Short-Time Fourier Transform (STFT) was performed to convert the time-domain signal into the frequency domain.
- **Mel Scale and Log Transformation**: The frequency spectrum was mapped to the Mel scale to simulate human auditory perception. A logarithmic transformation was then applied to the Mel-scaled energies.
- **Cepstral Coefficients Calculation**: The Discrete Cosine Transform (DCT) was used to derive 13 MFCCs from the log-Mel energies, producing a compact and informative representation of the spectral content.

The resulting MFCC matrix for each segment was flattened into a one-dimensional array using the `.flatten()` function, yielding a feature vector suitable for model input. In total, 10,000 MFCC-encoded feature vectors were obtained, each representing a 3-second audio segment.

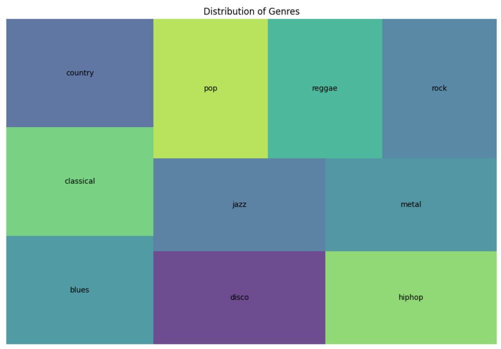


Figure 1: Balanced genre distribution in GTZAN: 100 tracks per genre ensure unbiased learning.

3 Model Architectures

3.1 CNN

A three-layer CNN is built with batch normalization, max pooling, and a dense dropout layer, concluding with a softmax output layer for classification.

3.2 RNN

The RNN uses stacked LSTM layers with dropout regularization to model temporal dependencies in the audio data.

3.3 SVM

An SVM with an RBF kernel is trained on flattened MFCC vectors, preceded by feature standardization using `StandardScaler`.

3.4 Random Forest

RF consists of 100 trees using square-root feature sampling at each split and handles non-linear decision boundaries well.

4 Evaluation Results

4.1 Without K-means Clustering

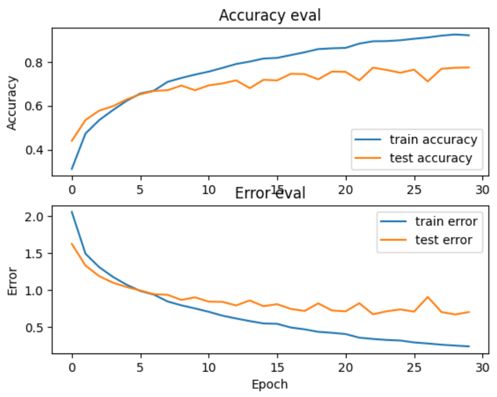
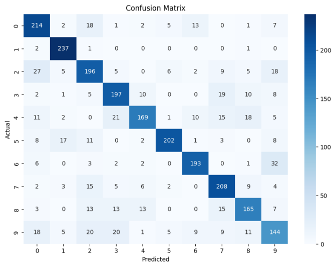


Figure 2: CNN accuracy across epochs: Fast convergence and high final accuracy.



	0	1	2	3	4	5	6	7	8	9
0	214	2	18	1	2	5	13	0	1	7
1	2	237	1	0	0	0	0	0	1	0
2	27	5	196	5	0	6	2	9	5	18
3	2	1	5	197	10	0	0	19	10	8
4	11	2	0	21	169	1	10	15	18	5
5	8	17	11	0	2	202	1	3	0	8
6	6	0	3	2	2	0	193	0	1	32
7	2	3	15	5	6	2	0	208	9	4
8	3	0	13	13	13	0	0	15	165	7
9	18	5	20	20	1	5	9	9	11	144

Figure 3: CNN confusion matrix: Strong performance in classical and metal; some confusion in pop.

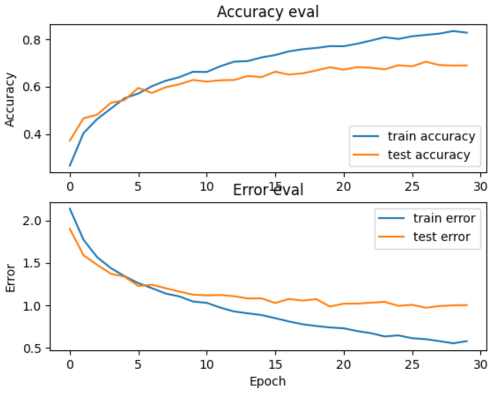
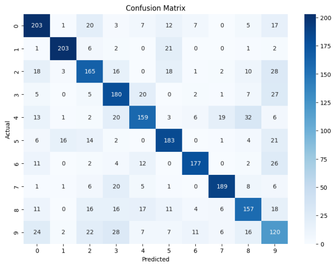


Figure 4: RNN accuracy shows less stability and lower final accuracy.



	0	1	2	3	4	5	6	7	8	9
0	203	1	20	3	7	12	7	0	5	17
1	1	203	6	2	0	21	0	0	1	2
2	18	3	165	16	0	18	1	2	10	28
3	5	0	5	180	20	0	2	1	7	27
4	13	1	2	20	159	3	6	19	32	6
5	6	16	14	2	0	183	0	1	4	21
6	11	0	2	4	12	0	177	0	2	26
7	1	1	6	20	5	1	0	189	8	6
8	11	0	16	16	17	11	4	6	157	18
9	24	2	22	28	7	7	11	6	16	120

Figure 5: RNN confusion matrix: Confusions mainly in similar genres like reggae and hip-hop.

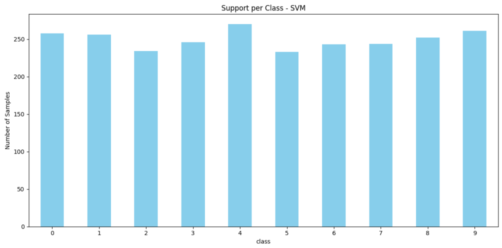
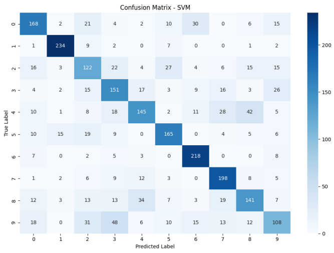
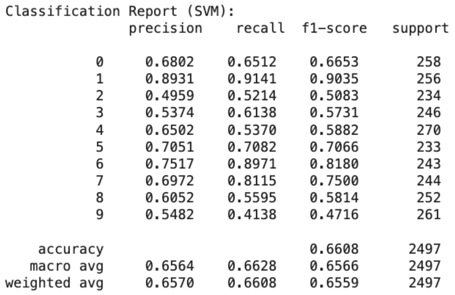


Figure 6: SVM support per genre: Indicates performance dependency on class characteristics.



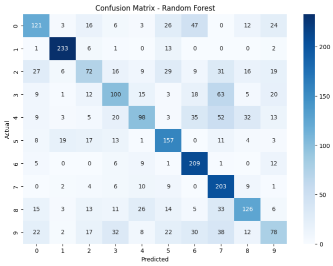
	0	1	2	3	4	5	6	7	8	9
0	168	2	21	4	2	10	30	0	6	15
1	1	234	9	2	0	7	0	0	1	2
2	16	3	122	22	4	27	4	6	15	15
3	4	2	15	151	17	3	9	16	3	26
4	10	1	8	18	145	2	11	28	42	5
5	10	15	19	9	0	165	0	4	5	6
6	7	0	2	5	3	0	218	0	0	8
7	1	2	6	9	12	3	0	198	8	5
8	12	3	13	13	34	7	3	19	141	7
9	18	0	31	48	6	10	15	13	12	108

Figure 7: SVM confusion matrix: Rock and pop are frequently confused.



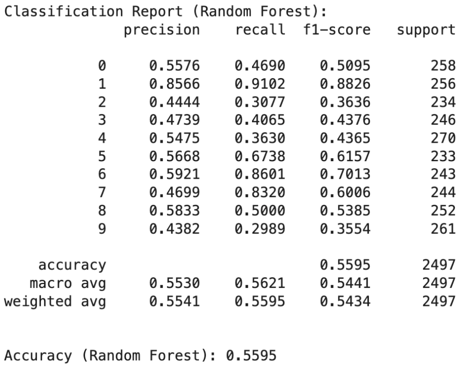
```
Classification Report (SVM):
              precision    recall  f1-score   support

           0     0.6802    0.6512    0.6653       258
           1     0.8931    0.9141    0.9035       256
           2     0.4959    0.5214    0.5083       234
           3     0.5374    0.6138    0.5731       246
           4     0.6502    0.5370    0.5882       270
           5     0.7051    0.7082    0.7066       233
           6     0.7517    0.8971    0.8180       243
           7     0.6972    0.8115    0.7500       244
           8     0.6052    0.5595    0.5814       252
           9     0.5482    0.4138    0.4716       261

    accuracy                         0.6608      2497
   macro avg     0.6564    0.6628    0.6566      2497
weighted avg     0.6570    0.6608    0.6559      2497
```

Figure 8: SVM classification metrics per genre. Accuracy varies by genre complexity.



	0	1	2	3	4	5	6	7	8	9
0	121	3	16	6	3	26	47	0	12	24
1	1	233	6	1	0	13	0	0	0	2
2	27	6	72	16	9	29	9	31	16	19
3	9	1	12	100	15	3	18	63	5	20
4	9	3	5	20	98	3	35	52	32	13
5	8	19	17	13	1	157	0	11	4	3
6	5	0	0	6	9	1	209	1	0	12
7	0	2	4	6	10	9	0	203	9	1
8	15	3	13	11	26	14	5	33	126	6
9	22	2	17	32	8	22	30	38	12	78

Figure 9: RF confusion matrix: Fewer errors than SVM, especially in jazz and blues.



```
Classification Report (Random Forest):
              precision    recall  f1-score   support

           0     0.5576    0.4690    0.5095       258
           1     0.8566    0.9102    0.8826       256
           2     0.4444    0.3077    0.3636       234
           3     0.4739    0.4065    0.4376       246
           4     0.5475    0.3630    0.4365       270
           5     0.5668    0.6738    0.6157       233
           6     0.5921    0.8601    0.7013       243
           7     0.4699    0.8320    0.6006       244
           8     0.5833    0.5000    0.5385       252
           9     0.4382    0.2989    0.3554       261

    accuracy                         0.5595      2497
   macro avg     0.5530    0.5621    0.5441      2497
weighted avg     0.5541    0.5595    0.5434      2497

Accuracy (Random Forest): 0.5595
```

Figure 10: RF classification report: Balanced performance across most genres.

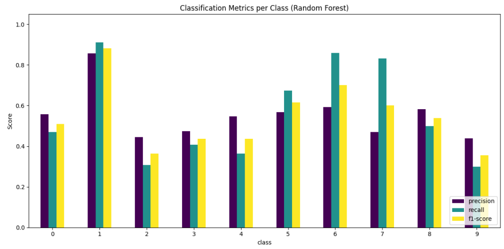
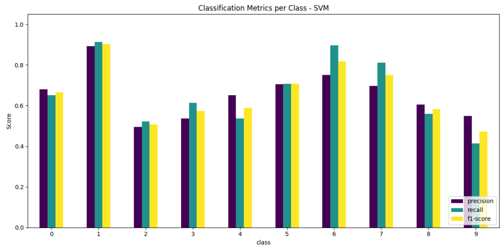


Figure 11: RF per-class accuracy: High for classical and metal; variable for others.



SVM per-class accuracy: High for classical and metal; variable for others.

4.2 With K-means Clustering

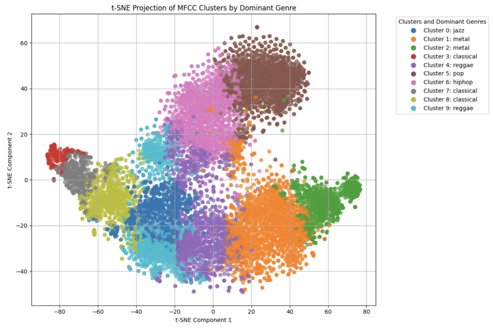


Figure 12: t-SNE plot of MFCC clusters post K-means. Misaligned clusters suggest added noise.

5 Model Comparison and Analysis

CNN outperforms all models in terms of test accuracy and training efficiency. RNN captures temporal patterns but is computationally heavier. SVM offers good interpretability but struggles with training time. RF provides a strong balance and robust performance.

6 Conclusion and Future Work

CNN is best suited for MFCC-based music genre classification. Clustering degrades performance by disrupting feature structures. Future research may explore hybrid CNN-RNN architectures, attention mechanisms, and multi-feature inputs. This paper presented a comparative analysis of CNN and RNN architectures units for music genre classification using MFCCs as input data. The results demonstrate that CNNs outperform RNNs in both accuracy and training efficiency, achieving higher classification accuracy and significantly faster training times. This advantage stems from CNNs' ability to extract spatial features from MFCC spectrograms, which naturally align with their convolutional structure. Additionally, the findings reveal that K-means clustering negatively impacts both CNN and RNN performance, as it fails to capture spatial and temporal dependencies essential for effective classification. This study offers valuable insights into the suitability of different approaches for audio pattern recognition. Future research could explore hybrid models that integrate CNN and RNN components, investigate alternative clustering techniques, and leverage larger, more diverse datasets to further enhance model performance and generalization.

References

- Course materials: `https://www.unimi.it/en/education/degree-programme-courses/2025/audio-pattern-recognition`

- GTZAN Dataset: `https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification`

- Tzanetakis, G., and Cook, P. (2002). Musical genre classification of audio signals. *IEEE Trans. Speech Audio Process.*, 10(5), 293–302.

- Humphrey, E. J., Bello, J. P., and LeCun, Y. (2013). Feature learning and deep architectures. *J. Intell. Inf. Syst.*, 41(3), 461–481.