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

Moein Ghaeini H.
Matriculation Number: 14460A
MSc Computer Science, University of Milan
Prof. Stavros Ntalampiras

June 4, 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 plays a key role in recommendation systems, music retrieval, and digital media organization. This work examines the performance of deep learning methods (CNN, RNN) and traditional machine learning classifiers (SVM, RF) using MFCCs extracted from the GTZAN dataset. Additionally, the impact of unsupervised clustering via K-means is analyzed.

2 Dataset and Feature Extraction

2.1 GTZAN Genre Collection

The GTZAN dataset consists of 1000 audio tracks equally distributed across 10 genres. Each 30-second track is sampled at 22050 Hz and split into 3-second segments to generate 10,000 samples. The 10 genres include: Blues, Classical, Country, Disco, Hip-hop, Jazz, Metal, Pop, Reggae, and Rock.

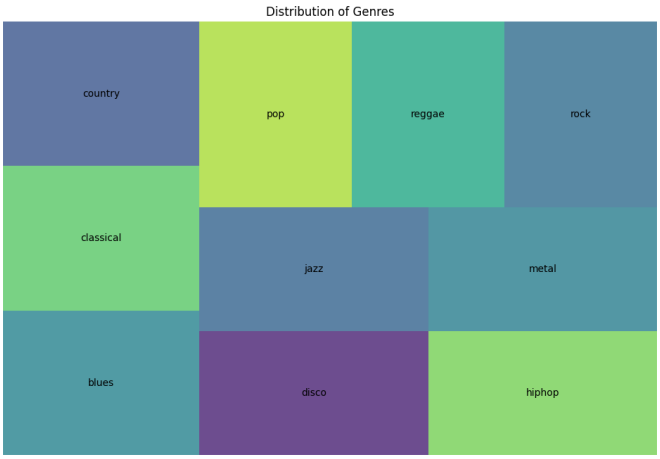

Figure 1: Balanced genre distribution in GTZAN: 100 tracks per genre ensure unbiased learning.

2.2 MFCC Extraction

MFCCs are extracted using 2048-sample FFTs with a hop length of 512, generating 13 coefficients per frame. Each audio segment yields a time-series MFCC matrix.

3 Data Processing Pipeline

3.1 Audio Preprocessing and Segmentation

Each audio track is normalized and segmented into 3-second clips. The MFCCs are then extracted from these segments.

3.2 Optional Augmentation

Augmentation techniques include pitch shifting, time stretching, and Gaussian noise addition to enhance dataset variability and model generalization.

4 Model Architectures

4.1 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.

4.2 RNN

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

4.3 SVM

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

4.4 Random Forest

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

5 Evaluation Results

5.1 Without K-means Clustering

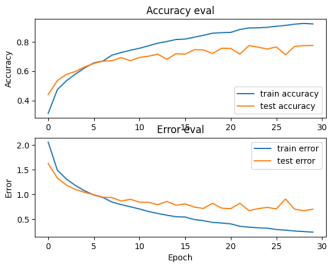

Figure 2: CNN accuracy across epochs: Fast convergence and high final accuracy.

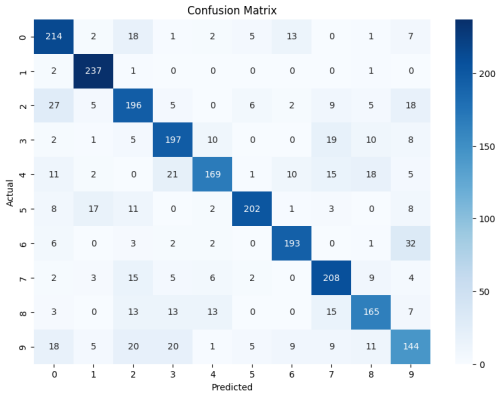

Figure 3: CNN confusion matrix: Strong performance in classical and metal; some confusion in pop.

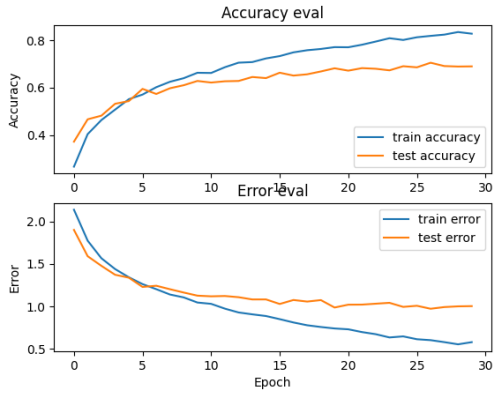

Figure 4: RNN accuracy shows less stability and lower final accuracy.

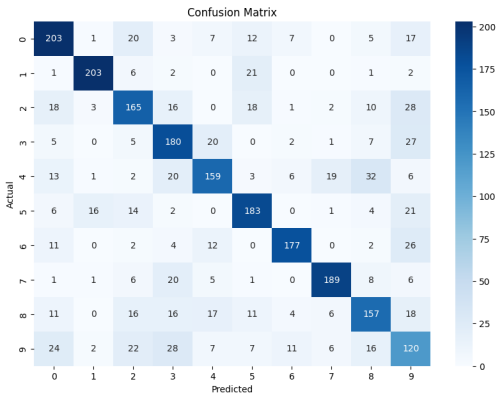

Figure 5: RNN confusion matrix: Confusions mainly in similar genres like reggae and hip-hop.

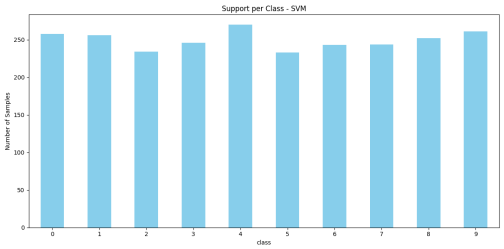

Figure 6: SVM support per genre: Indicates performance dependency on class characteristics.

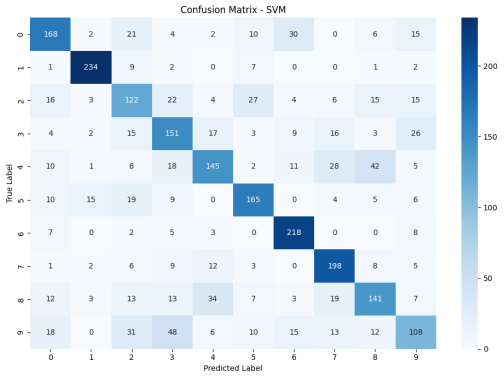

Figure 7: SVM confusion matrix: Rock and pop are frequently confused.

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.

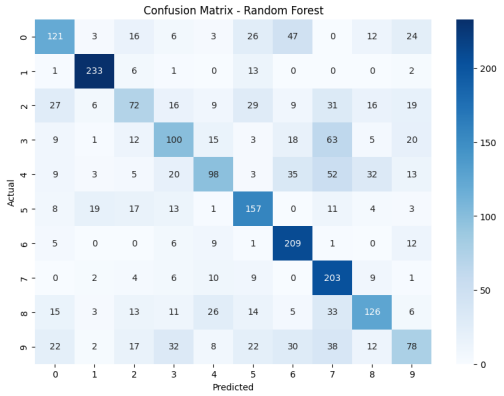

Figure 9: RF confusion matrix: Fewer errors than SVM, especially in jazz and blues.

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.

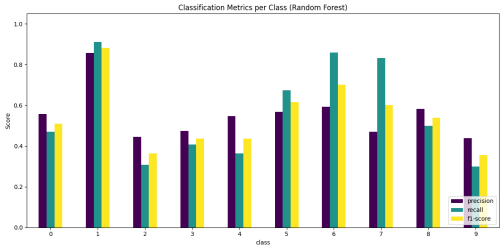

Figure 11: RF per-class accuracy: High for classical and metal; variable for others.

5.2 With K-means Clustering

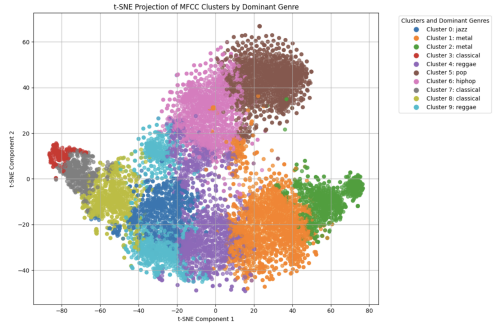

Figure 12: t-SNE plot of MFCC clusters post K-means. Misaligned clusters suggest added noise.

6 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.

Prof. Stavros Ntalampiras

7 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.

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.