Audio Genre Classification Using XGBoost and Feature Engineering

Tomer Teperovich & Lukov Snir Meir
M.Sc. in Machine Learning and Data Science
Reichman University (IDC). aka The Interdisciplinary Center
Herzliya, Israel
tomerteper2@gmail.com

Abstract

Signal Processing for Artificial Intelligence in Audio is an interdisciplinary field that combines traditional signal processing techniques with modern AI algorithms to analyze, interpret, and generate audio data. This project focuses on audio style classification, categorizing audio samples into predefined genres. Using the GTZAN dataset, various machine learning approaches were explored, ultimately achieving 93% accuracy with XGBoost. This paper presents the dataset, feature extraction methods, experimental setup, results, and future directions.

1 Introduction

Audio genre classification aims to categorize audio samples into predefined genres based on their spectral and temporal characteristics. This task is challenging due to variations in tempo, instrumentation, and recording quality. Traditional deep learning models such as CNNs have struggled with this task due to data limitations and model complexity. In this work, we propose an alternative approach using structured feature extraction and XGBoost, a gradient boosting decision tree algorithm, which achieves high classification accuracy with improved efficiency.

2 Dataset and Feature Extraction

The GTZAN dataset [1] is used for training and evaluation. It consists of 1000 audio tracks spanning ten genres: Blues, Classical, Country, Disco, Hip-Hop, Jazz, Metal, Pop, Reggae, and Rock. Each track is a 30-second waveform sampled at 22,050 Hz. Understanding and Preprocessing Audio Data GTZAN ipynb

2.1 Feature Extraction

To facilitate effective classification, we extract the following features:

- **Mel-Frequency Cepstral Coefficients (MFCCs)** - Capture spectral characteristics of audio.
- **Chroma Features** - Represent tonal content through pitch intensity analysis.
- **Spectral Contrast** - Measures differences between spectral peaks and valleys.
- **Spectral Centroid** - Represents the brightness of sound.

Feature visualization techniques, including time-domain representations, mel spectrograms, and chroma feature plots, were used to analyze the dataset.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

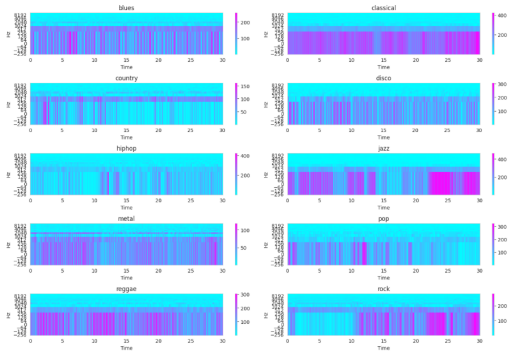

Figure 1: MFCC

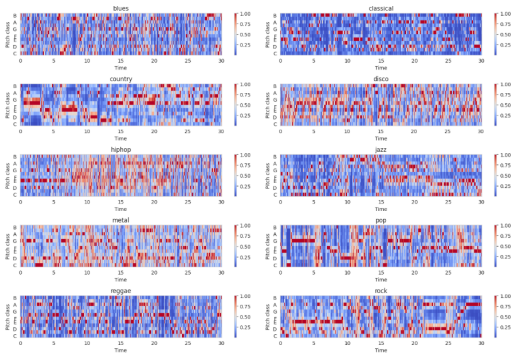

Figure 2: Chroma

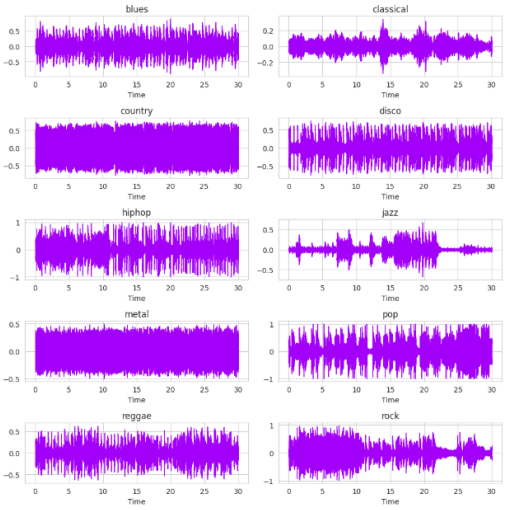

Figure 3: Wave

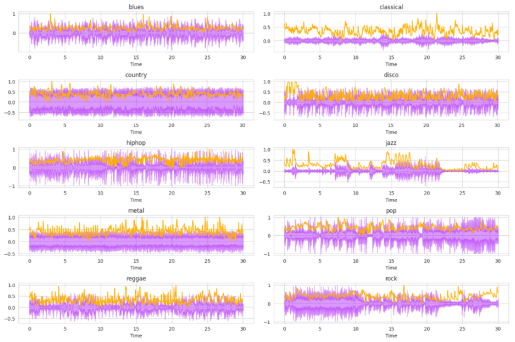

Figure 4: Spectral Centroids

3 Methodology

We evaluated different models before selecting the optimal approach:

3.1 Multi-Layer Perceptron

MLP: The initial model tested was a Multi-Layer Perceptron (MLP), which primarily relied on extracted feature vectors but struggled to capture the rich temporal and spectral variations in audio data, resulting in low classification accuracy 59.00% and poor generalization. MLP VS RF ipynb

3.2 Deep Learning Models

CNNs: Convolutional Neural Networks trained on spectrogram images showed limited success, achieving a maximum accuracy of 64% due to data constraints and the inability to model long-term dependencies effectively.CNN ipynb

3.3 XGBoost Approach

Due to the limitations of deep learning models, we employed XGBoost, which offers:

- Robustness to noise.
- Efficient feature selection.
- Strong performance on small-to-medium datasets.

Instead of using raw spectrogram images, we extracted MFCCs, spectral contrast, and chroma features as structured inputs for XGBoost achieving an accuracy of 71% XGB Classifier 71 Grid Seearch CV ipynb. Additionally, we segmented the audio into 3-second clips before feature extraction, improving performance by reducing intra-class variability, which led to achieving an accuracy of 93%. XGB Classifier 93 ipynb

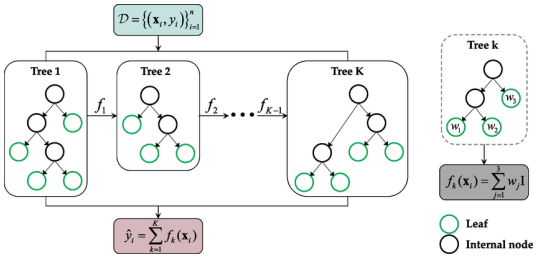

Figure 5: XGBoost Diagram

4 Experiments and Results

4.1 Preprocessing

Feature extraction was performed using MFCCs, chroma features, and spectral contrast. Data augmentation techniques such as time-stretching and pitch-shifting were applied to enhance model generalization.

4.2 Training Setup

The XGBoost model was optimized through hyperparameter tuning, adjusting:

- Number of estimators
- Maximum depth
- Learning rate

A 90-10 train-test split was used, and cross-validation ensured robustness. The final XGBoost model achieved **93% accuracy**, significantly outperforming CNN-based approaches.

Genre	Precision	Recall	F1-Score
Blues	0.95	0.96	0.96
Classical	0.96	1.00	0.98
Country	0.88	0.94	0.91
Disco	0.95	0.92	0.93
Hip-Hop	0.95	0.90	0.92
Jazz	0.93	0.94	0.93
Metal	0.92	0.97	0.95
Pop	0.98	0.92	0.95
Reggae	0.92	0.92	0.92
Rock	0.89	0.85	0.87

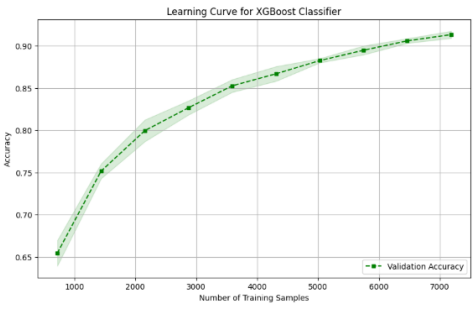

Figure 6: Learning Curve XGBoost

5 Future Work

Several areas can be explored to improve classification performance:

- **Hybrid Models:** Ensemble learning techniques integrating multiple models.
- **Advanced Feature Engineering:** Additional features such as tempo and zero-crossing rate.
- **Transfer Learning:** Utilizing pre-trained models and fine-tuning them.
- **L3-Like Embeddings:** Leveraging OpenL3 embeddings to enhance feature representation. XGB Classifier openl3 ipynb
- **Real-Time Applications:** Optimizing for low-latency streaming inference.

6 Conclusion

This work demonstrates that XGBoost can outperform traditional deep learning models in music genre classification by leveraging structured feature extraction. The proposed approach achieved 93% accuracy, significantly surpassing CNN-based methods. Future research should explore ensemble techniques and multimodal learning to further refine classification performance.

References

[1] Tzanetakis, G., and Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*.

[2] Doras, R., Cohen-Hadria, A., and Richard, G. (2021). Music Genre Classification: A Review of Deep-Learning and Traditional Machine-Learning Approaches. extitIEEE Access. Available: https://ieeexplore.ieee.org/document/9422487

[3] Cramer, J., Wu, H., Salamon, J., Bello, J. P., Ellis, D. P. (2019). Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings. extitIEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Available: https://arxiv.org/abs/1905.11787