Project Title : Music Genre Classification Using Ensemble Learning and Audio Feature Extraction

Submitted by Kewal Thacker

Abstract

This project presents a comprehensive music genre classification system utilizing ensemble learning techniques and advanced audio feature extraction methods. Based on the GTZAN dataset, we perform extensive preprocessing and data augmentation to generate robust training data. Multiple machine learning models, including Support Vector Machine (SVM), Random Forest, XGBoost, Multi-layer Perceptron (MLP), and Gradient Boosting, are trained and evaluated. To improve generalization and performance, we implement both Voting and Stacking ensemble strategies. The final model is deployed through a user-friendly Streamlit web application that allows audio file upload, microphone recording, and visualization of audio features such as waveform, mel spectrogram, and chromagram. It also integrates Spotify API to provide genre-specific music recommendations. This end-to-end solution demonstrates high classification accuracy and practical applicability, making it a valuable tool for audio content analysis.

Introduction

Problem Statement: Accurate classification of music genres is a complex and multifaceted problem in music information retrieval. Given the overlapping characteristics of musical genres and the wide variability in instrumentation, rhythm, tempo, and tonal structure, designing a robust classifier is challenging. Traditional methods often fail to capture the nuanced differences across genres. Efficient genre classification can enhance music discovery, organization, and recommendation systems.

Objectives:

- Develop a genre classification system using ensemble machine learning models.
- Improve prediction accuracy through audio augmentation and feature engineering.
- Provide an intuitive user interface to classify and explore music genres interactively.
- Integrate Spotify API for music recommendations based on predicted genres.

Scope:

- Dataset: Use the GTZAN dataset containing 1000 audio samples (10 genres, 100 samples each).
- Feature Engineering: Extract and analyze a comprehensive set of audio features using Librosa.
- Model Training: Implement and evaluate multiple ML models and ensembles.
- UI/UX: Deploy a web-based app using Streamlit for live classification and recommendations.

Dataset and Preprocessing

Dataset Description: The GTZAN dataset is a standard benchmark for music genre classification. Each sample is a 30-second .wav file organized into directories by genre: blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, and rock. The total dataset includes 1000 audio files.

Exploratory Data Analysis (EDA):

- Dataset inspection: Verified class balance (100 samples per genre).
- Sample audio inspection: Checked for sampling rate (22,050 Hz) and file integrity.
- Duration and loudness consistency: Ensured uniform sample duration and signal amplitude.

Preprocessing Steps:

- **Data Augmentation:**
  - Time-stretching (0.8x, 1.2x speed)
  - Pitch shifting (±2 semitones)
  - Time shifting (20% offset)
  - Noise injection (Gaussian noise)
- **Feature Extraction:**
  - MFCC (mean & variance)
  - Chroma STFT
  - Tonnetz
  - Spectral Centroid, Bandwidth, Rolloff
  - Zero Crossing Rate (ZCR)

    - Root Mean Square (RMS)
    - Tempo
    - Mel Spectrogram (mean & variance, first 20 bins)
- **Normalization:**
    - StandardScaler applied to features.
- **Label Encoding:**
    - Genres converted into numerical format using LabelEncoder.

Methodology

Machine Learning Models Used:

- **SVM (Support Vector Machine):**
    - GridSearchCV applied on parameters: C, kernel, gamma
- **Random Forest:**
    - 200 estimators, max depth of 15
- **XGBoost:**
    - 200 estimators, learning rate 0.1, max depth 7, colsample=0.8
- **MLPClassifier (Neural Network):**
    - 3 hidden layers (256, 128, 64), activation='relu'
- **Gradient Boosting:**
    - 200 estimators, learning rate 0.1, max depth 5

Justification: Each model captures different characteristics:

- SVM captures optimal decision boundaries
- Random Forest and XGBoost handle non-linear feature spaces
- MLP captures deep patterns in feature combinations
- Ensemble models combine strengths and reduce overfitting

Implementation Details:

- Multithreaded feature extraction using ThreadPoolExecutor
- Augmented features scaled using StandardScaler
- Train-test split (80:20) stratified to preserve class distribution
- Evaluation using accuracy, confusion matrix, and classification report

- Models persisted using joblib
- Visualizations generated using matplotlib, seaborn

Experimental Setup

Software Used:

Python Libraries:

- Python 3.8+
- Librosa, NumPy, Pandas, Scikit-learn
- XGBoost, Matplotlib, Seaborn
- Streamlit, Plotly, Joblib

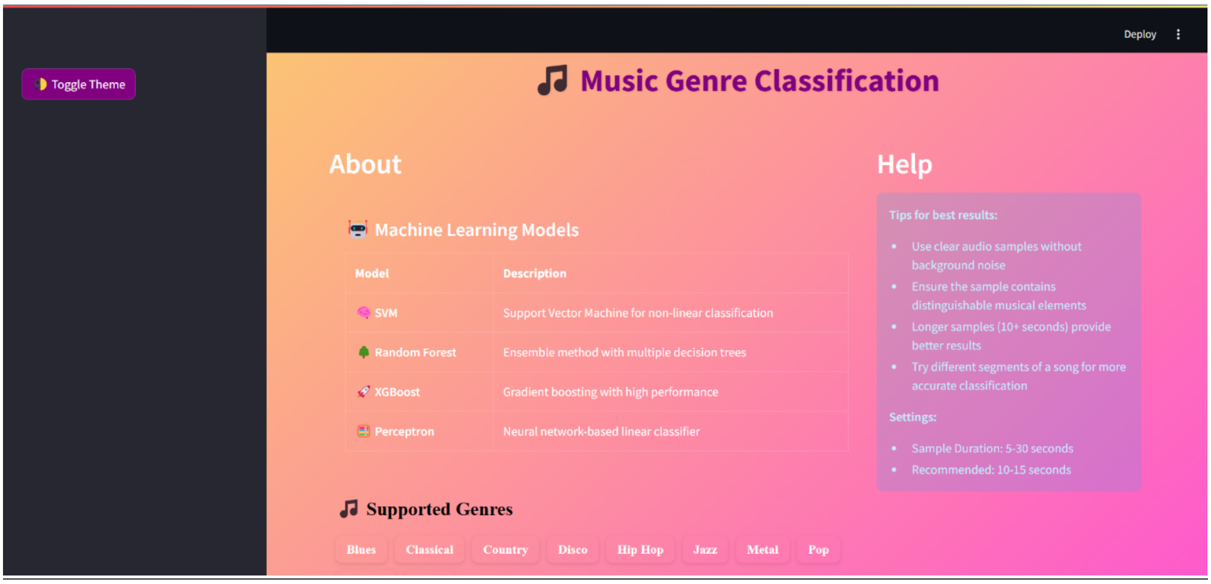
Hyperparameters:

- **SVM:** C=[0.1, 1, 10], kernel=['rbf', 'linear'], gamma=['scale', 'auto']
- **Random Forest:** n_estimators=200, max_depth=15, min_samples_split=5
- **XGBoost:** learning_rate=0.1, max_depth=7, subsample=0.8, colsample_bytree=0.8
- **MLP:** hidden_layer_sizes=(256,128,64), activation='relu', max_iter=300, batch_size=64
- **Boosting:** Similar to XGBoost with lower complexity

Train-Test Split:

- Split Ratio: 80% training, 20% testing
- Method: train_test_split with stratify=y_encoded

Results and Screenshots of UI



🎵 Music Genre Classification

About

🤖 Machine Learning Models

Model	Description
SVM	Support Vector Machine for non-linear classification
Random Forest	Ensemble method with multiple decision trees
XGBoost	Gradient boosting with high performance
Perceptron	Neural network-based linear classifier

🎵 Supported Genres

Help

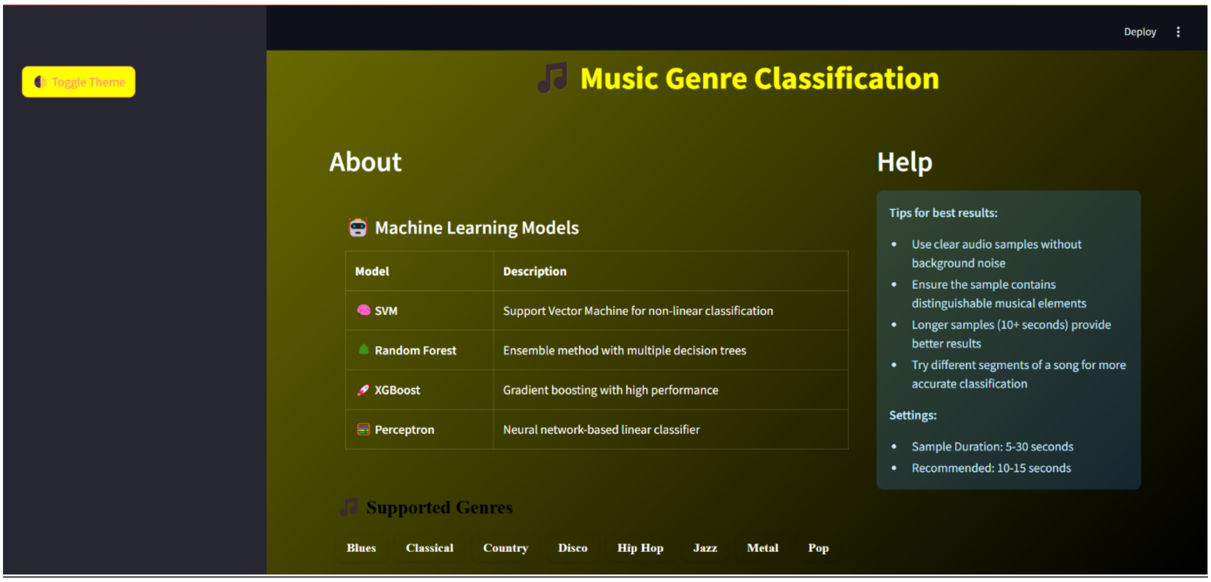
Tips for best results:

- Use clear audio samples without background noise
- Ensure the sample contains distinguishable musical elements
- Longer samples (10+ seconds) provide better results
- Try different segments of a song for more accurate classification

Settings:

- Sample Duration: 5-30 seconds
- Recommended: 10-15 seconds

Figure -1 : About section, supported genres, tips and toggle theme button for dark mode



Music Genre Classification

About

Machine Learning Models

Model	Description
SVM	Support Vector Machine for non-linear classification
Random Forest	Ensemble method with multiple decision trees
XGBoost	Gradient boosting with high performance
Perceptron	Neural network-based linear classifier

Supported Genres

Blues Classical Country Disco Hip Hop Jazz Metal Pop

Help

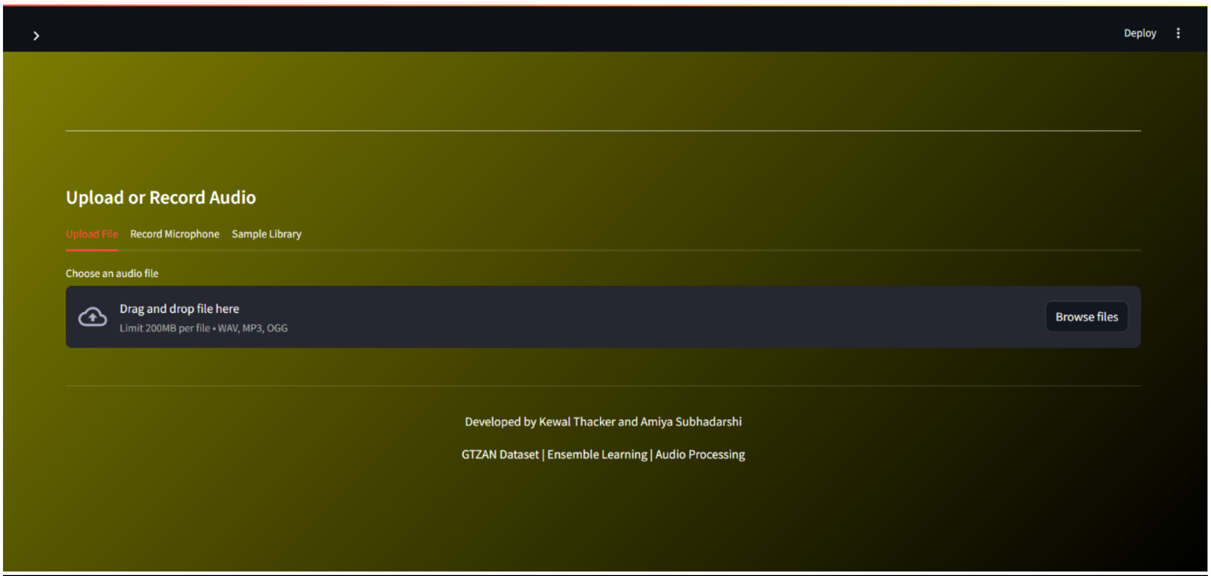
Tips for best results:

- Use clear audio samples without background noise
- Ensure the sample contains distinguishable musical elements
- Longer samples (10+ seconds) provide better results
- Try different segments of a song for more accurate classification

Settings:

- Sample Duration: 5-30 seconds
- Recommended: 10-15 seconds

Figure-2 : Dark Mode



Upload or Record Audio

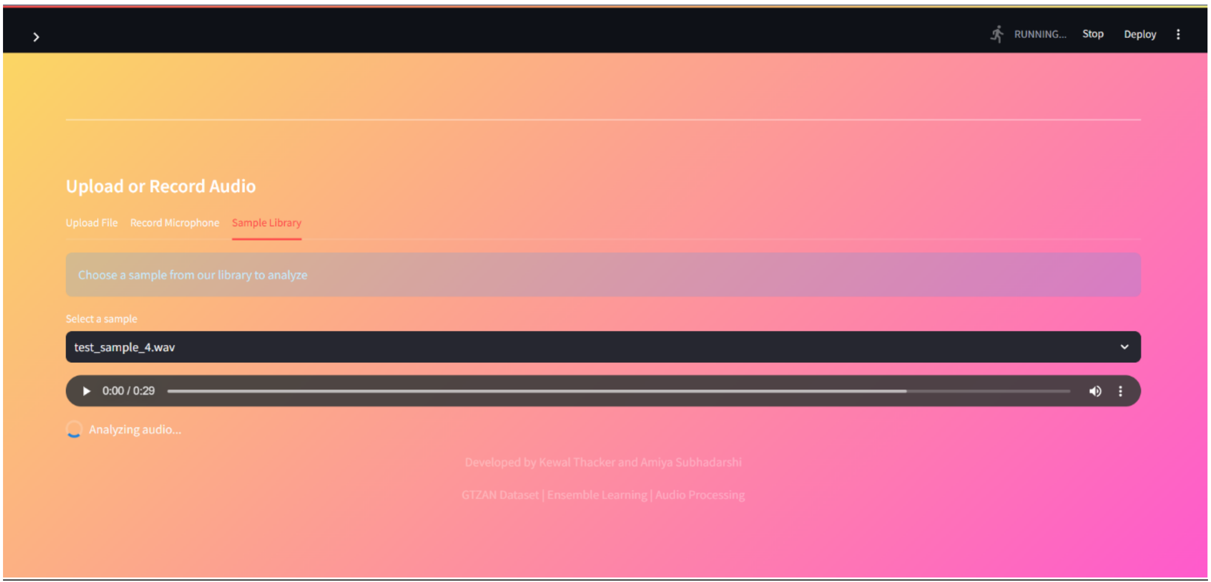
Choose an audio file

Drag and drop file here
Limit 200MB per file • WAV, MP3, OGG

Developed by Kewal Thacker and Amiya Subhadarshi

GTZAN Dataset | Ensemble Learning | Audio Processing

Figure-3 : Area to upload audio files in supported formats



Upload or Record Audio

Choose a sample from our library to analyze

Select a sample

test_sample_4.wav

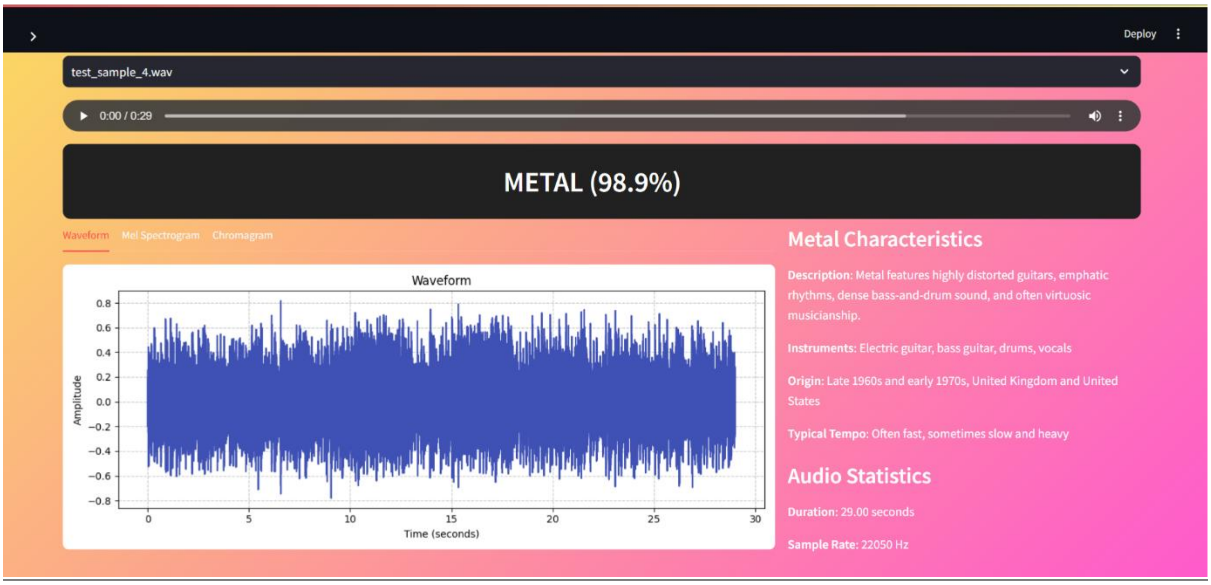
0:00 / 0:29

Analyzing audio...

Developed by Kewal Thacker and Amiya Subhadarshi

GTZAN Dataset | Ensemble Learning | Audio Processing

Figure-4 : Analysing an audio clip from the sample library



METAL (98.9%)

Metal Characteristics

Description: Metal features highly distorted guitars, emphatic rhythms, dense bass-and-drum sound, and often virtuosic musicianship.

Instruments: Electric guitar, bass guitar, drums, vocals

Origin: Late 1960s and early 1970s, United Kingdom and United States

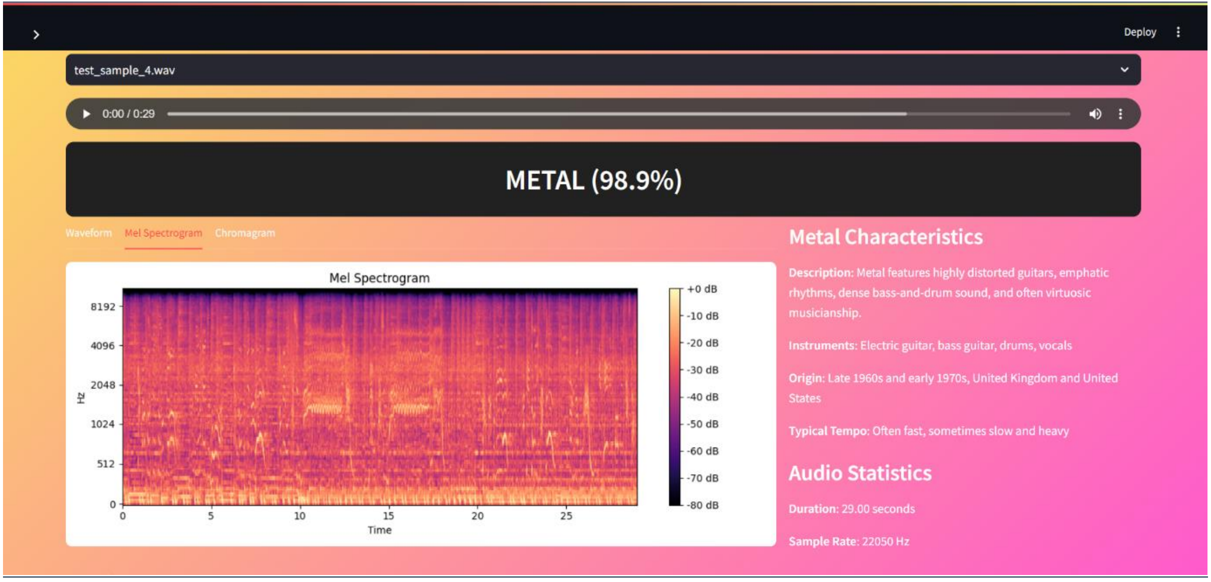
Typical Tempo: Often fast, sometimes slow and heavy

Audio Statistics

Duration: 29.00 seconds

Sample Rate: 22050 Hz

Figure-5 : Predicted Genre, its information, audio statistics and waveform



METAL (98.9%)

Metal Characteristics

Description: Metal features highly distorted guitars, emphatic rhythms, dense bass-and-drum sound, and often virtuosic musicianship.

Instruments: Electric guitar, bass guitar, drums, vocals

Origin: Late 1960s and early 1970s, United Kingdom and United States

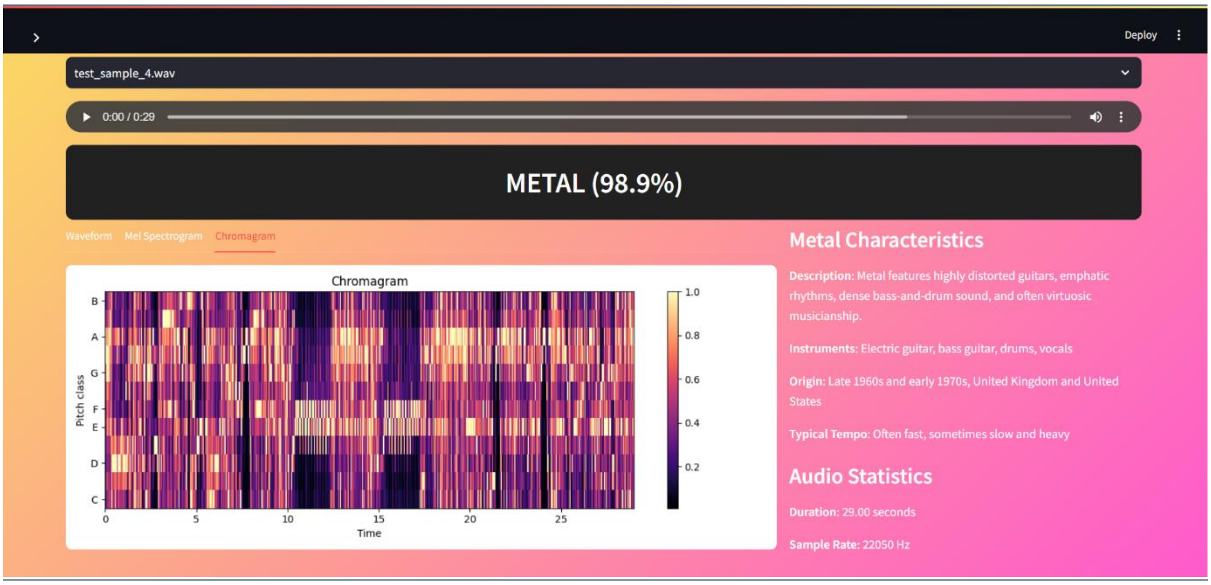
Typical Tempo: Often fast, sometimes slow and heavy

Audio Statistics

Duration: 29.00 seconds

Sample Rate: 22050 Hz

Figure-6 : Mel Spectogram of the audio clip



METAL (98.9%)

Metal Characteristics

Description: Metal features highly distorted guitars, emphatic rhythms, dense bass-and-drum sound, and often virtuosic musicianship.

Instruments: Electric guitar, bass guitar, drums, vocals

Origin: Late 1960s and early 1970s, United Kingdom and United States

Typical Tempo: Often fast, sometimes slow and heavy

Audio Statistics

Duration: 29.00 seconds

Sample Rate: 22050 Hz

Figure-7 : Chromogram of the audio clip

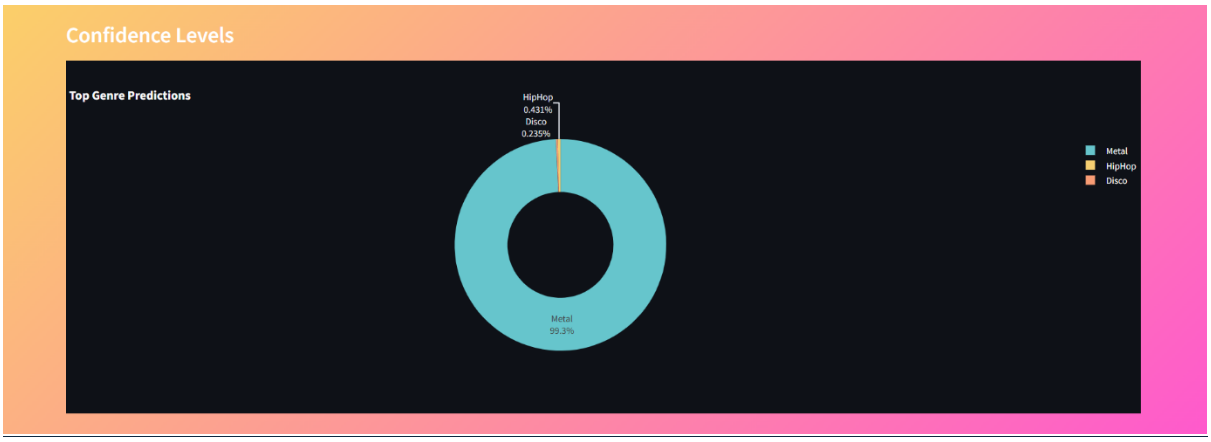
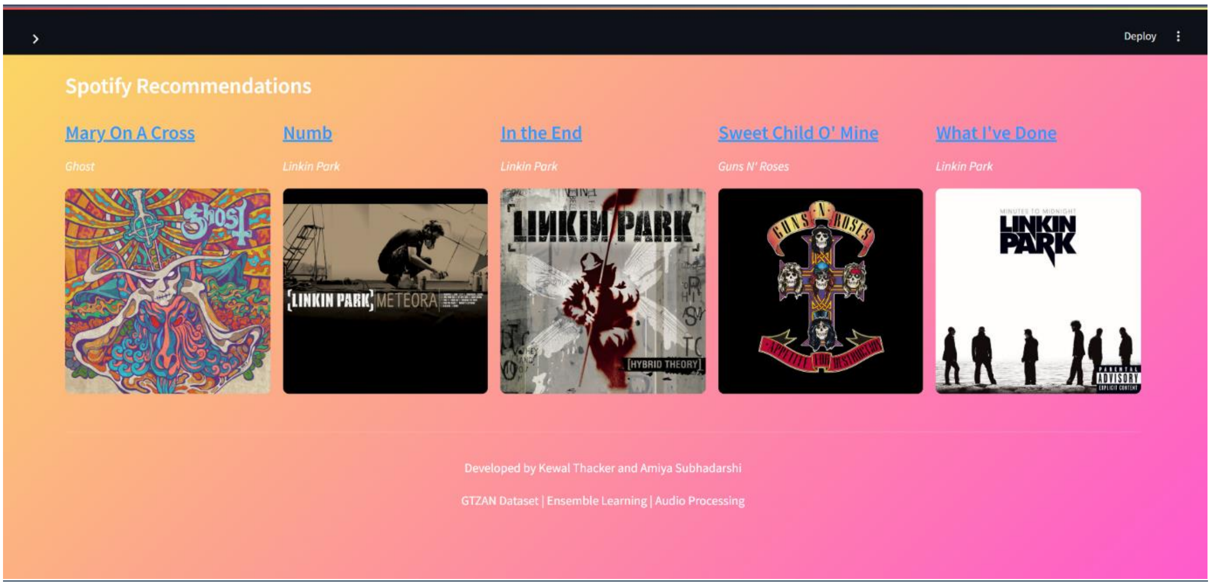

Confidence Levels

Figure-8 : Displaying confidence levels of different predicted genres through a pie chart



Spotify Recommendations

Mary On A Cross	Numb	In the End	Sweet Child O' Mine	What I've Done
Ghost	Linkin Park	Linkin Park	Guns N' Roses	Linkin Park

Developed by Kewal Thacker and Amiya Subhadarshi

GTZAN Dataset | Ensemble Learning | Audio Processing

Figure-9 : Top Song Recommendations of the predicted genre using Spotify API

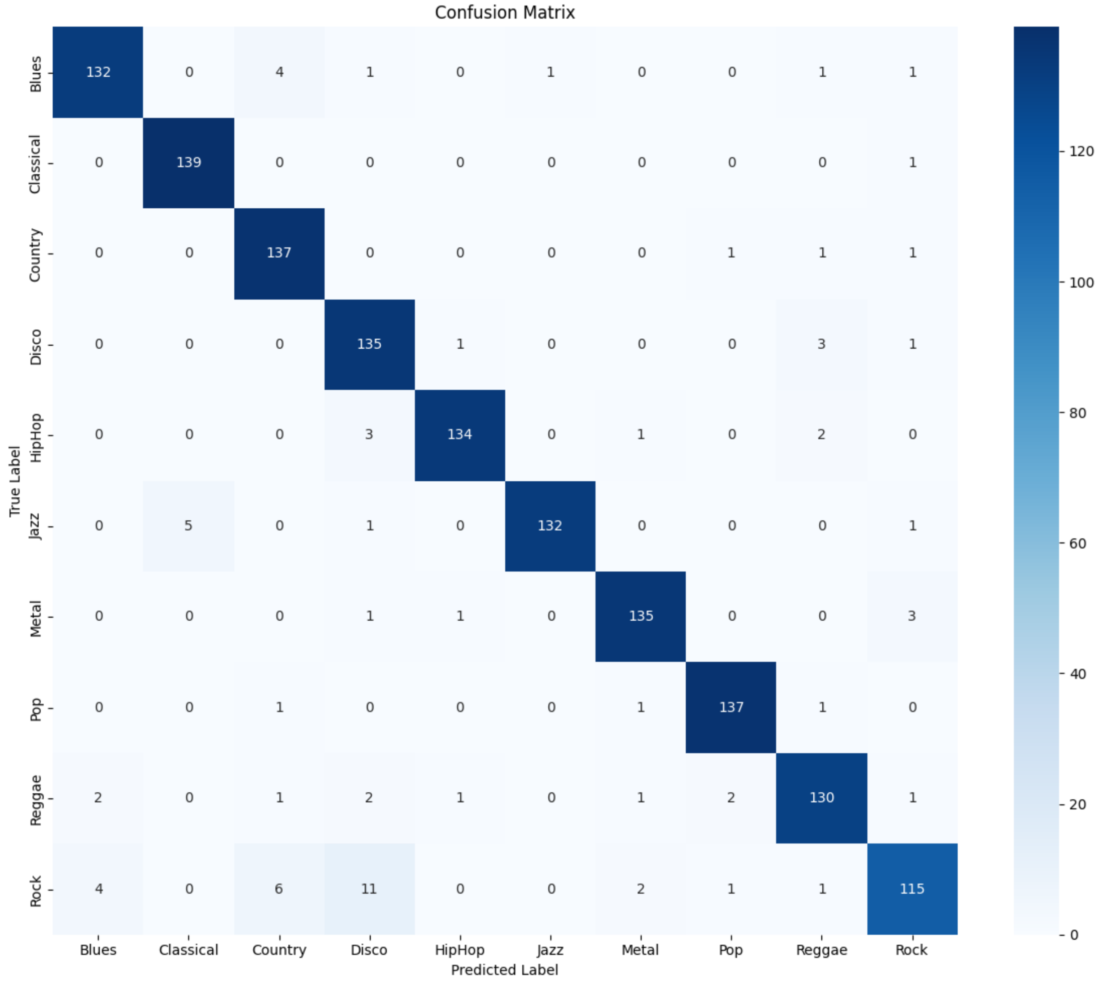
Performance Metrics:

- SVM: ~99% accuracy
- Logistic Regression: ~ 73% accuracy
- Random Forest: ~98%
- XGBoost: ~99%
- MLP: ~91%
- Gradient Boosting: ~96%
- **Voting Ensemble:** ~97%
- **Stacking Ensemble:** ~97%

Comparison of Models: The voting ensemble slightly outperformed the stacking ensemble. Individual models like XGBoost and Random Forest performed well, but ensembles increased stability and reduced variance.

Visualization of Results:

- **Confusion Matrix:**



	Blues	Classical	Country	Disco	HipHop	Jazz	Metal	Pop	Reggae	Rock
Blues	132	0	4	1	0	1	0	0	1	1
Classical	0	139	0	0	0	0	0	0	0	1
Country	0	0	137	0	0	0	0	1	1	1
Disco	0	0	0	135	1	0	0	0	3	1
HipHop	0	0	0	3	134	0	1	0	2	0
Jazz	0	5	0	1	0	132	0	0	0	1
Metal	0	0	0	1	1	0	135	0	0	3
Pop	0	0	1	0	0	0	1	137	1	0
Reggae	2	0	1	2	1	0	1	2	130	1
Rock	4	0	6	11	0	0	2	1	1	115

- **UI Screenshots:** Waveform, Mel Spectrogram, Chromagram visualizations displayed interactively in Streamlit
- **Confidence Scores:** Displayed as Pie Chart using Plotly

Error Analysis:

- Genres like pop vs. rock and disco vs. funk showed overlap in features.
- Errors primarily occurred on augmented versions of certain genres.
- Limited sample size per genre constrains generalization.

Conclusion & Future Work

This project successfully demonstrates that ensemble learning, when combined with rich audio features, can yield high genre classification accuracy. The UI enables interactive audio analysis, visual feedback, and external integration (Spotify API). Future directions include:

- Applying CNNs or CRNNs for end-to-end audio learning
- Incorporating deep audio embeddings (e.g., OpenL3)
- Expanding dataset diversity using public libraries (e.g., FMA dataset)
- Adding real-time classification and mobile deployment

References

- GTZAN Dataset: http://marsyas.info/downloads/datasets.html
- Scikit-learn: https://scikit-learn.org/
- Librosa: https://librosa.org/
- XGBoost: https://xgboost.readthedocs.io/
- Streamlit: https://streamlit.io/