Audio-Based Music Classification

Subodh Kant
M23CSA531
CSE, IITJ

Syam Krishnan Sakthidharan
M23CSA535
CSE, IITJ

Abstract

Music classification and Recommendation has received much attention from MIR researchers in recent years. In the MIR community, an annual event Music Information Retrieval Evaluation eXchange (MIREX) is held for competitions on important tasks in MIR since 2004. Most of the high-level tasks in MIREX competitions are relevant to music classification. This project focuses on two core tasks: Genre Classification and Mood Classification.

1. Introduction

A key problem in MIR is classification and Recommendation, which assigns labels to each song based on genre, mood, etc. Music classification is an interesting topic with many potential applications. It provides important functionalities for music retrieval, as most end users may only be interested in certain types of music. Thus, a classification system would enable them to search for the music they are interested in. It has wide applications in platforms like Spotify, Google Play, and Apple Music.

To implement this, one of the most important steps is to classify the genre and mood of a music track. This requires audio processing, one of the most complex tasks that involves time signal processing, time series, spectrograms, spectral coefficients, and audio feature extraction to feed a neural network.

2. Related Work

Previous research in music classification and Recommendation has included traditional ML techniques such as SVM, Random Forest, and k-NN. With the advancement of deep learning, CNNs have shown great potential in learning high-level features from audio spectrograms. Research in mood classification, especially with the DEAM dataset, often incorporates both audio and emotion features to improve classification accuracy.

3. Methodology

3.1. Music Genre Classification Model

1. Feature Extraction: Two types of features are extracted to suit different modeling approaches
   - For Traditional Machine Learning: MFCC Features - Mel-Frequency Cepstral Coefficients (MFCCs) are extracted using librosa.feature.mfcc with 13 coefficients. Additional Features: Spectral centroid and chroma features are computed using librosa.feature.spectral-centroid and librosa.feature.chroma-stft.
   - For Convolutional Neural Network (CNN): Mel Spectrogram - Computed using librosa.feature.melspectrogram with 128 Mel bands. Converted to decibel scale (librosa.power-to-db) and normalized to [0, 1]. Spectrograms are standardized to 128x128 by truncating longer sequences or padding shorter ones with zeros.
2. Data Preprocessing:
   - Dataset Splitting: The data (MFCC features, spectrograms, and labels) is split into training (80) and testing (20) sets using train-test-split with stratification to maintain genre balance
   - Scaling (Traditional ML): MFCC features are standardized using StandardScaler to ensure zero mean and unit variance, improving model convergence.
   - Reshaping (CNN): Spectrograms are reshaped to include a channel dimension (128x128x1) for compatibility with the CNN input layer.
3. Model Development: We implemented ML models to classify music genres using the GTZAN dataset. The models include:
   - Support Vector Machines (SVM)
   - Random Forest
   - k-Nearest Neighbors (k-NN)
   - Convolutional Neural Network (CNN)
4. Model Evaluation
   - Accuracy: Proportion of correctly classified samples.
   - Classification Report: Precision, recall, and F1-score per genre.

- Confusion Matrix: Visualized using Seaborn heatmaps to assess per-class performance.

5. Model Selection and Saving:
   - Comparison: Accuracy scores of all models (SVM, Random Forest, k-NN, CNN) are compared.
   - Best Model: The model with the highest test accuracy is selected.

3.2. Mood Classification Model

1. Data Preprocessing:
   - Traditional ML Features: Audio files are loaded using Librosa with a sampling rate of 22,050 Hz. Features extracted include - 13 MFCCs (Mel-frequency cepstral coefficients), averaged over time. Tempo, derived from onset strength. Spectral centroid, averaged over time.
   - Convolutional Neural Network (CNN): Mel spectrograms are generated from audio files using Librosa (n-mels=128). Spectrograms are reshaped to include a channel dimension ($128 \times 128 \times 1$) for CNN input.
2. Data Splitting
   - The dataset is split into training (80) and testing (20) sets using train-test-split with a random state of 42 for reproducibility.
   - Separate splits are maintained for ML features (X-ml, y-ml) and CNN spectrograms (X-spectrograms, y-numeric).
3. Model Development:
   - Random Forest (Traditional ML): A Random Forest Classifier is implemented with hyperparameter tuning using GridSearchCV. Parameters tuned: number of estimators (100, 200) and maximum depth (10, 20, None). Cross-validation (5-fold) is used to identify the best model configuration.
   - Convolutional Neural Network (CNN): Architecture: Two Conv2D layers (32 and 64 filters, 3×3 kernels, ReLU activation). MaxPooling2D layers (2×2) after each convolution. Flatten layer followed by a Dense layer (128 units, ReLU). Output layer (4 units, softmax activation). Compiled with Adam optimizer and sparse categorical crossentropy loss.
4. Model Training:
   - Random Forest: Trained on the extracted feature set (MFCCs, tempo, spectral centroid) using the training split.
   - CNN: Trained for 20 epochs with a batch size of 32. 20 percent of training data is used for validation. A ModelCheckpoint callback saves the model with the highest validation accuracy.
5. Model Saving and Evaluation:
   - Random Forest: The best estimator from GridSearchCV is saved as a .joblib file using joblib.dump.
   - CNN: The best-performing model (based on validation accuracy) is saved in Keras format (.keras) using ModelCheckpoint.
   - Metrics: Accuracy percentage. F1-score, precision, and recall via classification report. Confusion matrix for each model.

4. Results

4.1. Genre Classification Model

- SVM: 70.50% **(Best)**

```
=== SVM ===
Best Params: {'C': 10, 'kernel': 'rbf'}
Accuracy: 68.00%
Classification Report:
              precision    recall  f1-score   support

       blues       0.57      0.60      0.59        20
   classical       0.83      0.95      0.88        20
     country       0.64      0.70      0.67        20
       disco       0.61      0.70      0.65        20
      hiphop       0.48      0.55      0.51        20
        jazz       0.76      0.95      0.84        20
       metal       0.89      0.85      0.87        20
         pop       0.93      0.65      0.76        20
      reggae       0.52      0.55      0.54        20
        rock       0.67      0.30      0.41        20

    accuracy                           0.68       200
   macro avg       0.69      0.68      0.67       200
weighted avg       0.69      0.68      0.67       200
```

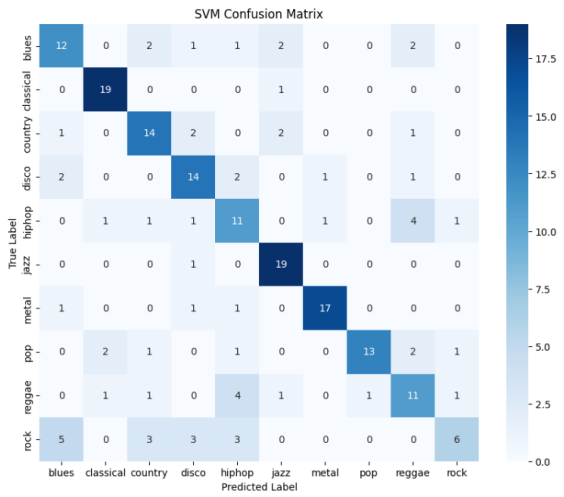

Figure 1. SVM Result - Genre Classification

- Random Forest: 66.50%

```
=== Random Forest ===
Best Params: {'max_depth': 10, 'n_estimators': 100}
Accuracy: 63.50%
Classification Report:
              precision    recall  f1-score   support

       blues       0.56      0.45      0.50        20
   classical       0.90      0.95      0.93        20
     country       0.54      0.65      0.59        20
       disco       0.57      0.80      0.67        20
     hiphop       0.38      0.40      0.39        20
        jazz       0.76      0.95      0.84        20
       metal       0.80      0.80      0.80        20
         pop       0.88      0.70      0.78        20
      reggae       0.48      0.50      0.49        20
        rock       0.38      0.15      0.21        20

    accuracy                           0.64       200
   macro avg       0.62      0.64      0.62       200
weighted avg       0.62      0.64      0.62       200
```

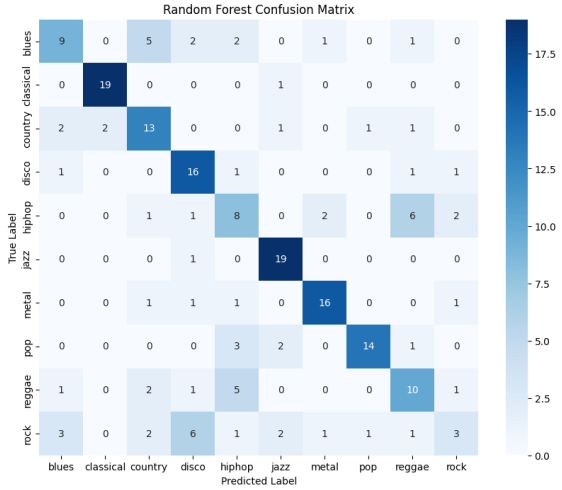

Figure 2. Random Forest Result - Genre Classification

- k-NN: 64.50%

```
=== k-NN ===
Best Params: {'n_neighbors': 7}
Accuracy: 63.50%
Classification Report:
              precision    recall  f1-score   support

       blues       0.56      0.45      0.50        20
   classical       0.94      0.80      0.86        20
     country       0.54      0.70      0.61        20
       disco       0.58      0.70      0.64        20
     hiphop       0.46      0.60      0.52        20
        jazz       0.76      0.95      0.84        20
       metal       0.89      0.80      0.84        20
         pop       0.79      0.75      0.77        20
      reggae       0.44      0.35      0.39        20
        rock       0.38      0.25      0.30        20

    accuracy                           0.64       200
   macro avg       0.63      0.64      0.63       200
weighted avg       0.63      0.64      0.63       200
```

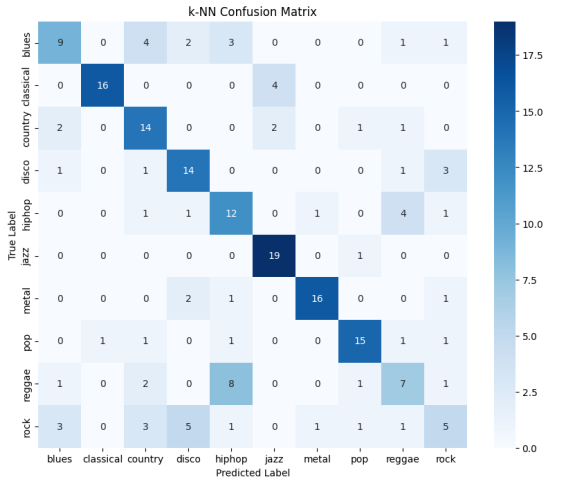

Figure 3. KNN Result - Genre Classification

- CNN: 61.50%

```
=== CNN ===
Accuracy: 50.00%
Classification Report:
              precision    recall  f1-score   support

       blues       0.38      0.25      0.30        20
   classical       0.89      0.85      0.87        20
     country       0.42      0.75      0.54        20
       disco       0.33      0.10      0.15        20
     hiphop       0.44      0.40      0.42        20
        jazz       0.62      0.50      0.56        20
       metal       0.69      0.90      0.78        20
         pop       0.45      0.45      0.45        20
      reggae       0.41      0.65      0.50        20
        rock       0.21      0.15      0.18        20

    accuracy                           0.50       200
   macro avg       0.49      0.50      0.48       200
weighted avg       0.49      0.50      0.48       200
```

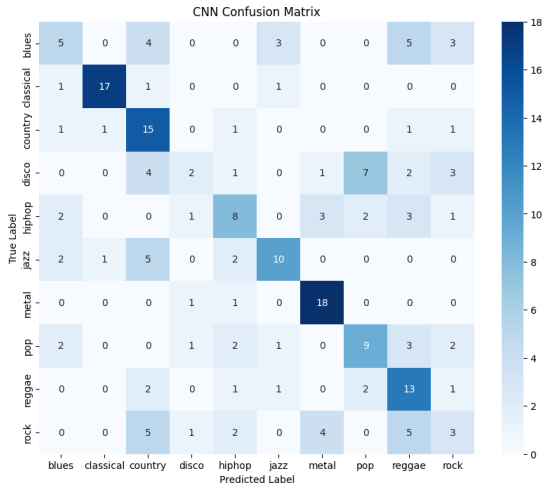

Figure 4. CNN Result - Genre Classification

4.2. Mood Classification Model

- Random Forest: **59.31% (Best)**

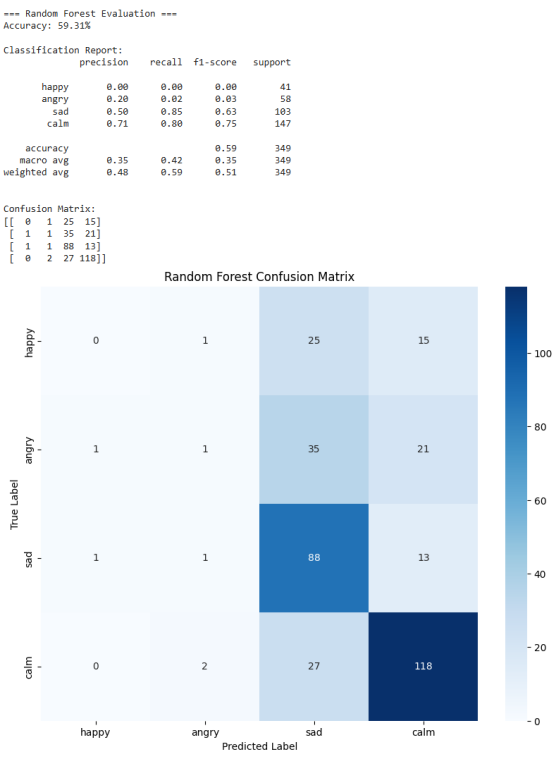

Figure 5. Random Forest Result - Mood Classification

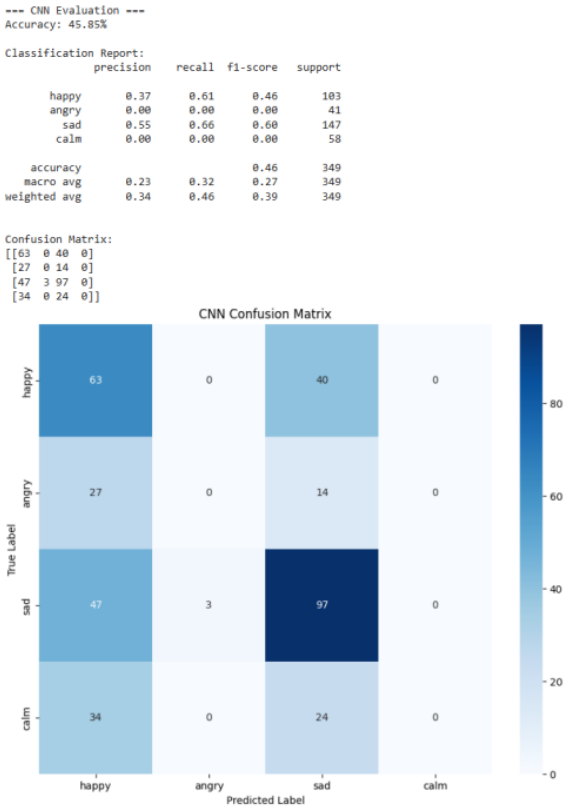

Figure 6. CNN Result - Mood Classification

- CNN: 45.85%

5. Analysis

- For genre classification, The best model was the SVM with 70.50% accuracy, outperforming others significantly.

```
=== Results Summary ===
SVM: 68.00%
Random Forest: 63.50%
k-NN: 63.50%
CNN: 50.00%

Best Model: SVM with accuracy 68.00%
```

Figure 7. Genre Classification - Result Summary

- For mood classification, the Random Forest model provided the best results with 59.31% accuracy.

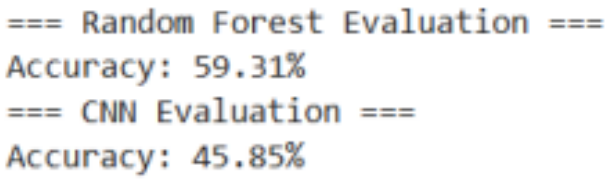

Figure 8. Mood Classification - Result Summary

- CNN models underperformed due to limited dataset size or architecture optimization issues.

6. Conclusion

This project aimed at automatic music genre and mood classification using deep learning and traditional ML models.

Phase A: The SVM provided the best performance at 70.50%.

Phase B: Random Forest achieved the highest mood classification accuracy at 59.31%.

References

[1] https://ieeexplore.ieee.org/document/5664797

[2] https://deepai.org/documents/P42iZ27

[3] https://www.researchgate.net/publication/221292434_Multi-level_Music_Mood_Classification_using_Audio_and_Lyrics

[4] https://www.sciencedirect.com/science/article/abs/pii/S0952197620315797?via%3Dihub