Music Genre Classification : Chloe Gray, Serene Qasem, Jordan Carden

Introduction

Music plays a central role in human culture and entertainment, and classifying music into genres is an essential task for organizing and exploring vast, diverse music collections. Especially considering the increasing size of digital music libraries, manual categorization is impractical, creating the need for an automated solution. Genre classification helps streamline the organization of music libraries, improve music recommendation systems, and improve user interactions on streaming platforms.

The rapid growth of digital music collections has made it increasingly difficult to manage, organize, and access content efficiently. Automating genre classification can save significant time and resources, while ensuring consistent and scalable results. Accurate genre classification can also enhance user recommendation systems, by delivering personalized playlists and improve user satisfaction.

Our group's project focuses on automating music genre classification, in which we tried different machine learning techniques to try to find the most accurate results possible. Using the GTZAN dataset from Kaggle, which contains audio samples across 10 genres, we evaluated two main approaches: CNN - Convolutional Neural Networks for image-based analysis of spectrograms, and Random Forest Classifiers, which excels in structured audio data classification by aggregating diverse tree predictions for robust and accurate genre identification.

Methods

The task of music genre classification involves assigning predefined genre labels to audio tracks based on their underlying features. This is a challenging problem due to the complexity of audio signals, which encompass a mix of pitch, tempo, rhythm, and more audio characteristics. Moreover, some genres share similar and overlapping attributes, making it difficult to distinguish between them.

The dataset used for this project contains 1,000 audio files spanning 10 genres, and each file is 30 seconds long. Features have been extracted for both 3 second and 30 second segments. These features include a variety of acoustic parameters, including spectral centroid, zero-crossing rate, and MFCCs. The dataset also includes spectrogram representations, which provide a visual depiction of frequency and amplitude variations over time.

Our project explores two machine learning approaches to solve the classification problem: the CNN model, which analyzes spectrogram images to extract patterns visually associated with different genres, and a Random Forest classifier, which uses structured numerical features derived from the audio files. Our goal is to evaluate the strengths and limitations of each method in terms of accuracy and scalability, ultimately contributing to the development of more effective music categorization systems.

The CNN model was designed to process spectrogram images, which visually represent the frequency content of audio signals over time. We preprocessed the spectrograms by converting them to grayscale and removing unnecessary white borders to provide the model with the most clean, concise images. This step ensured that the CNN model could better capture the

spectral and temporal patterns associated with each music genre. The network architecture was built on convolutional layers for feature extraction and fully connected layers for classification. The model was trained over 30 epochs, with early stopping applied if validation accuracy did not improve for five consecutive epochs.

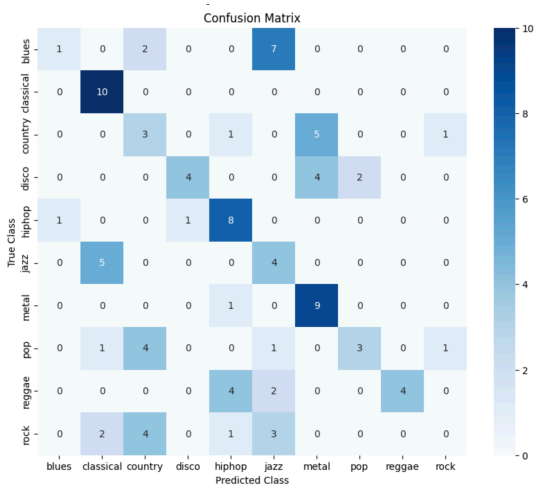
The second approach utilized a Random Forest classifier, which relied on numerical audio features such as spectral centroid, tempo, and MFCCs - Mel-frequency cepstral coefficients. These features, extracted in two different rounds with 3-second and 30-second audio segments, provided a structured dataset suitable for classification. The Random Forest model trained multiple decision trees, each using a random subset of the data and features to introduce diversity.

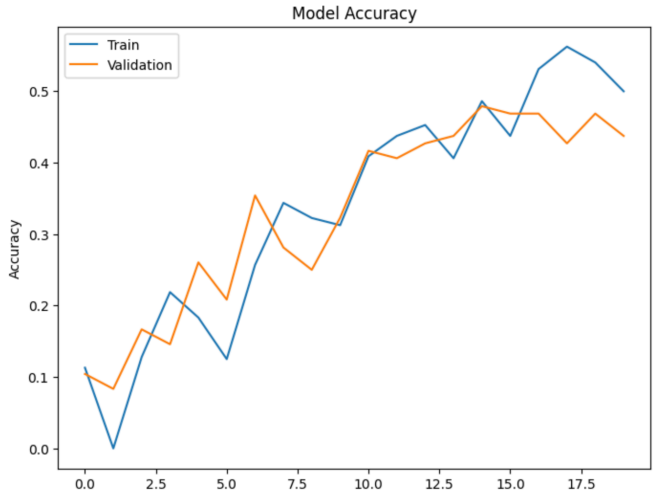
Results

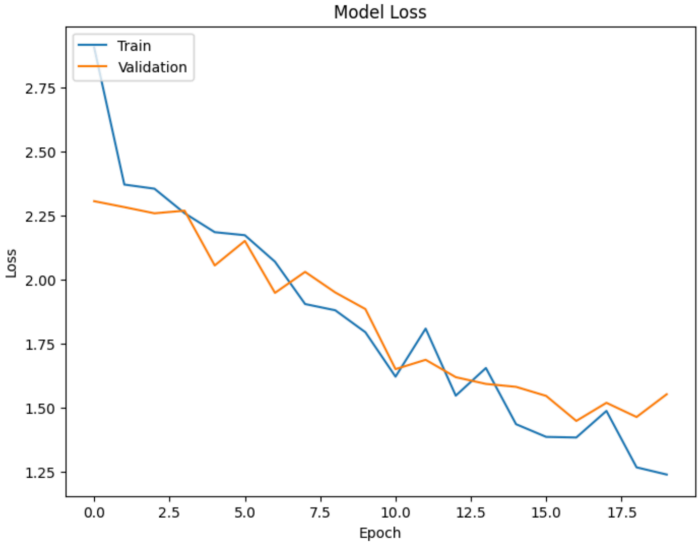
The CNN model achieved an accuracy of 55%, demonstrating its ability to identify genre-specific patterns to some extent. As shown in the confusion matrix below, it performed best for genres like Classical, Hip-Hop, and Metal, showing a promising ability to capture visual features in spectrogram images. However, the model struggled with genres like Rock and Blues which were frequently misclassified, due to challenges with distinguishing between overlapping audio features in these genres.

The training and validation accuracy and loss graphs for the CNN model, shown below, provide more details about its performance. The training accuracy graph shows a steady increase over the course of 30 epochs; however, the validation accuracy peaks around the 10th epoch, plateauing around 55% before declining. This divergence between training and validation

accuracy provides clear evidence of overfitting. The training loss graph shows a continuous decline, but the validation loss begins to rise after the 10th epoch. This reinforces the overfitting trend found within the model. Extending the training to 200 epochs or applying stronger regularization techniques, such as dropout or weight decay, could provide additional clarity on this overfitting trend and improve performance.







Classification Report

	precision	recall	f1-score	support
blues	0.50	0.10	0.17	10
classical	0.56	1.00	0.71	10
country	0.23	0.30	0.26	10
disco	0.80	0.40	0.53	10
hiphop	0.53	0.80	0.64	10
jazz	0.24	0.44	0.31	9
metal	0.50	0.90	0.64	10
pop	0.60	0.30	0.40	10
reggae	1.00	0.40	0.57	10
rock	0.00	0.00	0.00	10
accuracy			0.46	99
macro avg	0.50	0.46	0.42	99
weighted avg	0.50	0.46	0.42	99

The Random Forest classifier achieved a remarkable accuracy of 89%, significantly outperforming the CNN model's accuracy of 55%. By aggregating the predictions from multiple decision trees through majority voting, the Random Forest classifier demonstrated its strength in handling structured numerical data, particularly excelling with the 3-second audio segments. Its robustness and ability to identify key features influencing genre classification highlighted the strengths of feature-based methods for structured data.

The classification report and confusion matrix for the Random Forest provide additional details on the model's results. The model performed exceptionally well for Classical and Metal genres, achieving high recall and precision, with F1-scores above 0.9. However, it did have some challenges with similar-sounding genres like Jazz and Disco which were frequently misclassified. The model also struggled with Reggae, which had precise predictions but suffered from low recall. The Random Forest confusion matrix provides us with insight to genre-specific challenges, further explaining the misclassification between jazz, disco, and reggae. These findings highlight the strengths of the Random Forest model in handling distinct features while

identifying areas where further refinement could improve differentiation between overlapping genres.

```
Confusion Matrix:
[[16  0  1  0  0  1  0  0  1  1]
 [ 0 19  0  0  0  1  0  0  0  0]
 [ 2  0 15  0  0  1  0  0  0  2]
 [ 0  1  1 12  4  0  2  0  0  0]
 [ 1  0  0  1 16  0  1  0  0  1]
 [ 1  2  0  0  0 17  0  0  0  0]
 [ 0  0  0  2  0  0 17  0  1  0]
 [ 0  0  0  0  1  0  0 18  1  0]
 [ 0  0  0  0  1  0  0  3 16  0]
 [ 2  0  2  1  1  2  0  0  0 12]]
```

3-Second Feature Model performed better.

```
Accuracy for 3-Second Features: 0.8884
Classification Report:
              precision    recall  f1-score   support

       blues       0.88      0.89      0.88       200
   classical       0.91      0.97      0.94       199
     country       0.84      0.87      0.86       199
       disco       0.86      0.88      0.87       200
      hiphop       0.93      0.90      0.92       200
        jazz       0.86      0.91      0.88       200
       metal       0.89      0.95      0.92       200
         pop       0.96      0.83      0.89       200
      reggae       0.87      0.91      0.89       200
        rock       0.90      0.78      0.83       200

    accuracy                           0.89      1998
   macro avg       0.89      0.89      0.89      1998
weighted avg       0.89      0.89      0.89      1998

Confusion Matrix:
[[177   1   4   2   0   4   5   0   6   1]
 [  0 193   1   0   0   4   0   0   1   0]
 [  9   1 174   2   0   6   0   2   4   1]
 [  0   0   5 175   2   2   5   0   4   7]
 [  0   1   2   4 180   2   4   3   4   0]
 [  4  10   4   0   0 182   0   0   0   0]
 [  1   0   0   0   4   1 191   0   1   2]
 [  0   1   4  11   4   1   0 167   6   6]
 [  4   3   2   2   3   2   1   2 181   0]
 [  6   1  12   8   0   8   8   0   2 155]]
```

File display

```
Accuracy for 30-Second Features: 0.7900
Classification Report:
              precision    recall  f1-score   support

       blues       0.73      0.80      0.76        20
   classical       0.86      0.95      0.90        20
     country       0.79      0.75      0.77        20
       disco       0.75      0.60      0.67        20
      hiphop       0.70      0.80      0.74        20
        jazz       0.77      0.85      0.81        20
       metal       0.85      0.85      0.85        20
         pop       0.86      0.90      0.88        20
      reggae       0.84      0.80      0.82        20
        rock       0.75      0.60      0.67        20

    accuracy                           0.79       200
   macro avg       0.79      0.79      0.79       200
weighted avg       0.79      0.79      0.79       200
```

Conclusion

Comparing the two approaches, the Random Forest classifier proved to be more effective for this dataset, likely due to its ability to handle numerical features efficiently. However, the CNN approach holds promise, especially with potential improvements in preprocessing and architectural design. A combination of these methods, where CNN-derived features are fed into a Random Forest classifier, could potentially offer a hybrid solution that leverages the strength of both techniques.

This project examined two distinct approaches to music genre classification: a Convolutional Neural Network (CNN) and a Random Forest classifier. The CNN used spectrogram images to analyze audio visually, while the Random Forest focused on numerical features extracted from the audio files. The Random Forest proved to be more effective, achieving an accuracy of 89% compared to the CNN's 55%. This highlights the strength of feature-based methods for structured data like audio characteristics. Despite the CNN's lower performance, it demonstrated potential for improvement with better preprocessing and more advanced architectures. Overall, the comparison revealed valuable insights, suggesting that combining the strengths of both methods could lead to even better results in future work.