Final Project Report
Music Genre Classification
DS 5220, Fall 2024

Yogita Bisht[*] and Daniel Uriel González Quezada

In this report, we investigate music genre classification using machine learning and deep learning techniques. Initially, an exploratory data analysis (EDA) was performed to understand the dataset's characteristics, followed by feature extraction techniques using audio signal processing libraries such as LibROSA. A convolutional neural network (CNN) model was implemented to classify audio genres, leveraging features like Mel-frequency cepstral coefficients (MFCCs), spectral centroids, and chroma. Additionally, classical machine learning models, including Random Forest and Support Vector Machines (SVM), were optimized using hyperparameter tuning to improve classification performance. The results demonstrate significant accuracy improvements through advanced feature engineering and model selection. Finally, preliminary experiments in music generation using Variational Autoencoders (VAEs) were conducted, showcasing potential future directions in synthesizing audio signals from latent features.

I. OVERVIEW

A. The problem

The problem addressed in this project is the automatic classification of music genres based on audio signals. With the exponential growth of digital music libraries and streaming platforms, manually categorizing music into genres has become impractical. The objective is to build a system capable of accurately identifying the genre of a song from its audio features, aiding in organizing large music databases and enhancing user experiences through personalized recommendations.

The motivation behind this problem lies in the increasing demand for efficient and automated music classification systems. Such systems reduce manual effort and errors in tagging music and enable more refined content discovery for users.

B. Why is this problem interesting

Music genre classification is a problem at the intersection of audio signal processing, machine learning, and art. It is an essential task for content-based music retrieval systems, which are widely used in streaming platforms like Spotify, YouTube, and Apple Music.

On a broader scale, solving this problem has societal implications, as it facilitates cultural preservation by categorizing and archiving vast amounts of audio data. Additionally, automated systems can identify trends in musical styles and preferences, which can assist music producers, researchers, and marketers.

[*] GitHub repository

C. Proposed approach

This project employs a combination of classical machine learning algorithms and deep learning models to classify music genres. Initially, audio features such as Mel-frequency cepstral coefficients (MFCCs), spectral centroids, and chroma features were extracted using the LibROSA library. These features serve as the input to various classifiers, including Random Forest, Support Vector Machines (SVM), and Convolutional Neural Networks (CNNs).

While traditional machine learning models were fine-tuned using techniques like RandomizedSearchCV to optimize hyperparameters, CNNs were designed to exploit spatial hierarchies in spectrogram-like representations of the audio features. The project also explored Variational Autoencoders (VAEs) for generating new audio features, paving the way for potential audio synthesis applications.

D. Rationale behind the approach

The chosen approach takes some inspiration from existing works in music genre classification, where feature-based methods and neural networks have been shown to perform well. Prior works have demonstrated the efficacy of MFCCs and other spectral features in representing audio signals for classification tasks.

Unlike purely feature-based approaches that rely on manually engineered inputs, the use of CNNs allows the model to learn higher-order feature representations directly from the data, potentially improving classification performance. Moreover, the exploration of VAEs for feature generation represents an innovative extension, as few studies have combined classification and generation tasks in the same pipeline.

E. Results and limitations

The results demonstrate that classical machine learning models achieve moderate accuracy, while deep learning models like CNNs show higher accuracy due to their ability to learn complex patterns in the data. However, the approach is limited by the quality and diversity of the dataset. Future work could focus on expanding the dataset, exploring transfer learning, and enhancing VAE-based feature generation.

II. EXPERIMENT SETUP

A. Dataset

The dataset used in this project is the GTZAN Music Genre Collection, a well-known benchmark for music genre classification. It consists of 1,000 audio tracks evenly distributed across 10 genres: Blues, Classical, Country, Disco, Hip Hop, Jazz, Metal, Pop, Reggae, and Rock. Each track is approximately 30 seconds long and sampled at 22,050 Hz.

The dataset also included some preprocessed audio features for 30 and 3 second segments of each audio track, such as

- Mel-frequency cepstral coefficients (MFCCs)

- Spectral centroid and bandwidth
- Chroma features
- Root mean square (RMS) energy
- Zero-crossing rate

Each audio file's features were aggregated into summary statistics (mean, standard deviation, min, max) to form a feature vector.

B. Implementation

The experiments involved two distinct approaches: classical machine learning and deep learning. Each approach was designed to utilize the dataset's features effectively while exploring their strengths and limitations.

For the classical machine learning models, we implemented both Random Forest and Support Vector Machine (SVM) classifiers. The Random Forest model was configured with an initial number of 100 decision trees, and its depth was optimized to a maximum of 30 using RandomizedSearchCV. The SVM classifier utilized a radial basis function (RBF) kernel, with the regularization parameter C fine-tuned through cross-validation to balance the model's complexity and performance. These models were trained on the processed features from the audio tracks. Each feature vector consisted of 39 elements, including summary statistics (mean, standard deviation, minimum, and maximum) of various audio features such as Mel-frequency cepstral coefficients (MFCCs), spectral centroid, chroma features, and root mean square (RMS) energy.

For the deep learning approach, we built a Convolutional Neural Network (CNN) model tailored for music genre classification. Instead of using features, this model directly processed spectrogram-like representations of the audio data, leveraging the spatial patterns within these representations.

Lastly, we incorporated a Variational Autoencoder (VAE) as an exploratory step to generate new audio feature vectors. These autoencoders comprised an encoder-decoder structure, where the encoder reduced the input MFCC features to a latent dimension of 32 through progressively smaller layers (128, 64, and 32 units). The decoder mirrored the encoder's structure to reconstruct the input from the latent representation. The VAE was trained using a mean squared error (MSE) loss function, focusing on reconstructing MFCC-based features accurately. This exploration aimed to examine the potential of synthesized features for augmenting the dataset or serving as inputs to classification models.

Overall, the implementation balanced simplicity and sophistication, tailoring each model to effectively exploit the dataset's characteristics while addressing the computational challenges associated with training on limited resources.

III. RESULTS

A. Primary results

For the exploratory data analysis (EDA), we visualized spectrograms to observe distinct frequency patterns across music genres and created a boxplot to compare the BPM (beats per minute) of 10 genres, highlighting rhythmic differences. A heatmap was also constructed to examine correlations among audio features, focusing on Mel-frequency cepstral coefficients (MFCCs) and other acoustic characteristics. These insights provide a detailed understanding of feature relationships, aiding in feature selection and further analysis.

The CNN model leverages the spatial information in spectrograms to differentiate genres based on patterns in frequency and time. We developed a CNN model using Keras, incorporating convolutional layers, pooling layers, batch normalization, and dense layers to classify music genres. The model was trained and evaluated, achieving a training accuracy of 92.71 percent and a testing accuracy of 77.10 percent. To further enhance performance, a more sophisticated architecture such as the VGG16 model could be employed, leveraging its deeper layers and pre-trained weights to capture more complex features and improve accuracy.

The Random Forest classifier achieved an overall accuracy of approximately 88% after hyperparameter tuning, with its best performance observed for genres like Classical and Metal, where precision and recall exceeded 90%. However, genres with overlapping acoustic features, such as Blues and Country, exhibited lower precision and recall, reducing the model's ability to distinctly classify them.

The Support Vector Machine classifier outperformed Random Forest with an accuracy of about 92%. However, training the SVM took longer than the RF. Preliminary experiments with the Variational Autoencoder (VAE) for feature generation showed potential but were less conclusive. While the VAE successfully reconstructed MFCC-based features with low reconstruction loss, the generated features did not significantly enhance classification accuracy when used in conjunction with other models. Nonetheless, this remains an area for further exploration, especially for augmenting datasets.

Supplementary results

For the Random Forest classifier, the number of estimators (100) and maximum depth (30) were selected after performing a randomized search over a predefined parameter grid. These values balanced model complexity and performance while keeping computation times manageable.

For the SVM, we used a radial basis function (RBF) kernel due to its ability to handle non-linear relationships within the data. The regularization parameter was tuned to 1.0 after testing a range of values, which provided the best trade-off between bias and variance. Additionally, the features were standardized using z-score normalization, as SVMs are sensitive to feature scaling.

The CNN architecture was iteratively refined by testing different layer configurations. The final structure with three convolutional layers, max-pooling, and two dense layers offered the best balance of accuracy and generalization. Dropout rates of 0.3 were chosen after observing overfitting in initial experiments without regularization. The Adam optimizer was used with a learning rate of 0.001, as it demonstrated faster convergence compared to alternatives like SGD.

For the VAE, the latent dimension was set to 32 taken from similar works. This dimension effectively captured the variability in the MFCC-based features without introducing excessive noise.

IV. DISCUSSION

The results of this study reveal the strengths and limitations of classical machine learning models in the task of music genre classification. The Random Forest model achieved an accuracy of 88%, and the Support Vector Machine (SVM) classifier slightly outperformed it with an accuracy of 92%. These results underscore the utility of handcrafted audio features, such as Mel-frequency cepstral coefficients (MFCCs) and spectral centroids, in capturing essential characteristics of different music genres. However, the challenges of separating similar genres, such as Blues and Country, highlight the limitations of feature engineering when acoustic overlaps are significant.

Compared to existing approaches, these results align well with studies that use traditional classifiers on the GTZAN dataset, where accuracies often range from 80% to 90%. The careful optimization of hyperparameters and feature scaling likely contributed to the competitive performance observed. However, the dependency on handcrafted features restricts the models' ability to capture more nuanced patterns in the audio data.

On the other hand, CNN implementation did not perform very well. Our main intuition for this is the selected architecture might not be enough for the complexity of the data. Future work could focus on using more sophisticated CNN architectures, such as VGG16 to improve test accuracy.

However, as found in state-of-the-art approaches, it is necesarry to use multi-label classification, as songs can actually span several genres.

The exploration of Variational Autoencoders (VAEs) for feature generation offered insights into their potential applications. Nevertheless, results were not successful. Our idea of why this approach did not work was that the model worked with aggregated statistics ocmputed over the entire duration of a song rather than the raw audio signals or time-resolved representations. This aggregation simplifies the data but removes critical temporal and spectral details that might be necessary for meaningful feature synthesis and improved classification.

V. CONCLUSION

This project demonstrated the potential of classical machine learning techniques for music genre classification, achieving competitive results using Random Forest and Support Vector Machine models. The handcrafted audio features proved useful for this task, though challenges remain in separating genres with overlapping acoustic profiles.

The exploration of Variational Autoencoders (VAEs) revealed the limitations of using aggregated song-level statistics as inputs for feature generation. This simplification, while computationally efficient, likely contributed to the generative model's limited impact on classification performance. Future work focusing on time-resolved features and more advanced generative models could significantly enhance the utility of this approach, paving the way for improved classification and novel applications in music analysis and synthesis.