Genre-Based Classification of Songs Using Deep Learning Models

Aengus Martin G. Donaire
Fordham University, CIS Dept.
New York, New York USA
agd5@fordham.edu

***Abstract*—This study explores music genre classification using machine learning algorithms, focusing on the performance of deep learning models with audio feature inputs. While prior research has addressed multi-label classification and label ambiguity, gaps remain in model comparison and real-world applicability. Dense neural network (DNN), recuurent neural network (RNN), and convulotional neural network (CNN) architectures are developed and evaluated to determine which offers the most accurate and efficient classification. The dataset used in this study consists of short 3-second audio clips transformed into Mel Frequency Cepstral Coefficients (MFFCs). DNN showed the weakest performance, revealing its difficulty in learning localized patterns in audio data. RNN performed slightly better, taking advantage of its ability to handle temporal sequences but was limited by the short input length. CNN delivered the most promising results, effectively capturing spatial hierarchies in the data. Applying regularization techniques such as dropout further improved performance by preventing overfitting and enhancing generalization. Overall, the study shows that regularized CNN models are best suited for short-clip music genre classification, with future improvements possible through deeper networks, longer audio clips, and data enhancement.**

***Index Terms*—machine learning, music genre classification, deep neural networks**

I. Introduction

The classification of music genres using machine learning is a complex and dynamic field of research. It aims to categorize audio files into predefined genres based on their acoustic characteristics. During early machine learning efforts, when textual metadata was often more accessible than raw audio, text-based music genre classification is the main important area and especially relevant before wide access to high-quality audio data. With the rise of machine learning, more objective and automated methods for the classification of music genres were developed.

Deep machine learning models have demonstrated significant potential in this area. In particular, convolutional neural networks (CNNs) have been effectively trained using time-frequency representations of audio signals thereby achieving impressive classification performance, especially in terms of their accuracy. A substantial amount of research focused on music genre classification models. In fact, genre classification models that have been developed depend on the quality of the provided labels as part of the training process [3]. However, even though these models are advanced, there remains a gap in real-world applications, where some data may not contain the necessary text labels required for genre classification.

Despite these advancements, there is still a need for further improvement in both the accuracy and efficiency of genre classification models. A major challenge lies in the vast diversity of musical styles and sub-genres, which adds complexity to the classification process [9]. Additionally, it might also be the intricate nature of audio signals that contain a vast amount of temporal and spectral details. In fact, audio signals are unstructured data which contain continuous, multidimensional, and highly variable components [1], making feature extraction and classification difficult. Thus, the aim of this paper is to establish which machine learning algorithms can perform more accurately using input data extracted as features from audio files instead of text labels.

II. Related Works

Categorizing songs according to their genre is a challenging task in the area of music information retrieval. Sanden and Zhang (2011) [10] address the complexity of multi-label music genre classification by proposing the use of ensemble learning techniques to enhance predictive performance. Recognizing that music tracks often span multiple genres, their study argued that traditional single-label classification models are insufficient and thus combined various base classifiers to form an ensemble system capable of handling multi-label outputs more effectively. It emphasizes the importance of using multiple learning perspectives to better reflect the multifaceted nature of music genre categorization. However, the study focuses primarily on the performance of the classifier and does not comprehensively explore the role of label ambiguity or subjective genre of boundaries, which are crucial aspects of music classification [2]. The feature representation of music data is relatively limited potentially constraining the system's capacity to capture nuanced musical characteristics.

Trohidis et al. (2011) [11] utilized multi-label classification techniques to classify music with emotional labels, recognizing that a single track can evoke multiple emotional responses simultaneously. Utilizing supervised learning methods such as support vector machines and k-Nearest neighbors, the authors employ the MIREX dataset and extract audio features using the MIRtoolbox, offering a structured and replicable framework for emotion-based classification. However, while their study effectively demonstrates the advantages of multi-label learning, it does not account for the inherent ambiguity in emotional perception— an aspect that has gained more

attention in recent work, such as Buisson et al. (2022) [3], who adopt label distribution learning to explicitly model such uncertainty.

With the recent success of deep neural networks, a number of studies apply these techniques to classifying or describing audio files, aiming to characterize them more accuratel and robustly. As mentioned, Buisson et al. (2022) [3] introduced a label distribution learning framework to model the ambiguity inherent in music genre classification, allowing the model to learn from soft, probabilistic label distributions rather than relying on rigid single-label annotations. This approach reflects a growing recognition that music classification tasks benefit from models capable of capturing the complex, overlapping nature of musical categories [4]. However, this method lacks a comparative evaluation of different deep learning architectures under consistent conditions, particularly in terms of their generalization performance and sensitivity to regularization techniques.

In this study, a systematic comparison of three deep learning models is conducted— namely, dense neural networks (DNN), convolutional neural networks (CNN) with and without regularization, and recurrent neural networks (RNN)—by first transforming the music data into Mel-frequency cepstral coefficients (MFCCs), which effectively capture the timbral and spectral properties of the audio, to evaluate and contrast their effectiveness in music genre classification tasks.

III. Experimental Setup and Methods

Audio data can be a challenging format of data to work with, as it is disorganized and there are many aspects that require detailed knowledge of digital signal processing. In this study, a dataset from GTZAN was used. The raw audio dataset is obtained from Kaggle and is organized into genre subfolders (blues, classical, country, pop, hiphop, etc). Each genre has its own folder that consists of one-hundred 30-sec samples of music. This study used a collection of 10 genres with 100 audio files each, all having a length of 30 seconds.

A. Exploratory Data Analysis

The spectrum of a waveform represents how the signal's energy is distributed across different frequencies. It displays the magnitude of the signal— often measured in decibels (dB)—as a function of frequency. Simply, it breaks down a complex waveform into its constituent frequency components, allowing us to see which frequencies are present and how strong they are. This is especially useful in fields like audio analysis, telecommunications, and signal processing, where understanding the frequency content of a signal is crucial for interpretation, filtering, or modification. The frequency spectrum shown in Figure 1 illustrates the distribution of the magnitude of an audio signal, measured in decibels (dB), in a logarithmic frequency range from 1 Hz to 10 kHz. The spectrum reveals a rise in energy starting around 10 Hz, peaking between 50 and 250 Hz, which indicates strong low-frequency components typical of bass or low-pitched instruments. The presence of several narrow spikes and dips

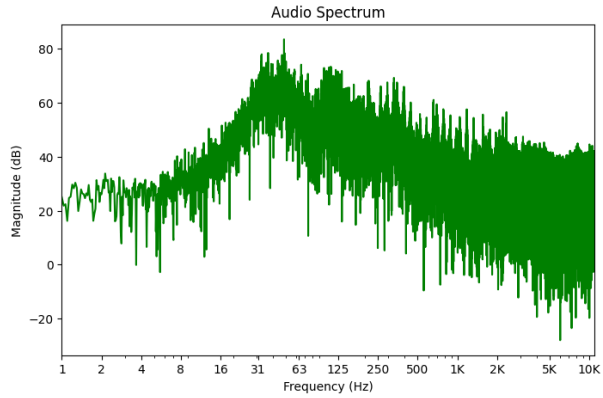

Fig. 1. Pop Music Spectrum

across the spectrum suggests noise, indicating that the signal is complex and likely consists of layered audio such as music, speech, or an environmental recording.

Although the spectrum provides a clear view of the frequency content, it does not account for how humans perceive sound, particularly at different frequencies. A Mel spectrogram addresses this by mapping the frequencies onto the Mel scale, which aligns more closely with human auditory perception, making it especially useful for tasks like speech and audio recognition. This is visualized using Mel Frequency Cepstral Coefficients (MFFCs) to obtain a more detailed and perceptually aligned representation of the audio's frequency content over time. Transforming audio into MFCCs enhances the data by emphasizing features that align with human auditory perception. This allows for significant dimensionality reduction while preserving essential information— effectively compressing the audio data.

B. Data Pre-processing

The collection process of MFFCs involves two key steps: (1) collecting audio samples and corresponding genres; and (2) converting audio to MFCCs. Collecting MFFCs from each audio file required dividing said files into 3-second segments to generate multiple samples from a single track. This segmentation is crucial for training deep learning models, as it increases the dataset size without requiring additional recordings. Each segment is then transformed into Mel-frequency cepstral coefficients (MFCCs), which capture the timbral and spectral characteristics of the audio. By generating multiple MFCC representations per song, the model is exposed to more varied inputs, improving its ability to learn features better and generalize across genres.

In this study, MFCCs were extracted using a sampling rate of 22,500 Hz, a standard choice that balances audio quality with computational efficiency. Each 30-second track was divided into ten 3-second segments, ensuring consistent temporal resolution across the dataset. The MFCCs were computed using a Fast Fourier Transform (FFT) window size of 2048 and a hop length of 512 samples, providing a detailed yet tractable spectral representation over time. Morevover, 13 MFCCs were

extracted per frame, capturing essential timbral characteristics while maintaining a low-dimensional feature space suitable for input to deep learning models. Only segments producing the expected number of feature vectors were retained, ensuring uniformity in input shape for model development.

In Figure 2, Mel-scaled spectrograms from different 3-second audio segments across various genres are displayed. Clear variations can be observed in the spectral energy distributions, particularly in the lower frequency bands, which often contain genre-specific rhythm. These differences highlight the relevance of time-frequency representations in capturing the unique characteristics of each genre.

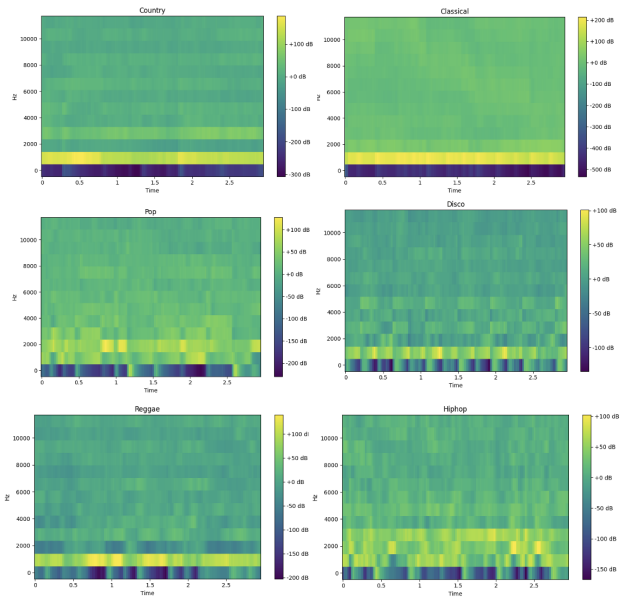

Fig. 2. Sample spectrograms by genre

C. Model Development

Figure 2 illustrates distinct spectro-temporal patterns across audio signals from different music genres, suggesting the presence of specific features for each genre. These characteristics, through MFCCs, serve as the input for model architectures to perform genre classification tasks. Specifically, the MFCC features and genre labels were extracted and split using stratified sampling: 70% of the data was initially allocated for training, with 30% held out as the test set. The training set was further split, allocating 30% of it for validation. This stratification ensured that the genre distributions remained consistent across all subsets.

Dense Neural Network DNN is composed of multiple dense layers ordered sequentially one after the other [6]. As shown in Figure 3, the implemented DNN model begins with a flatten layer that transforms the two-dimensional MFCC input of shape (130, 13) (i.e. 130 represents time steps and 13 represents MFCC coefficients) into a one-dimensional array of 1,690 features, preparing it for processing by dense layers. The network consists of three fully connected layers with 512, 256, and 64 neurons, respectively. Each dense layer is followed by normalization, a ReLU activation function, and a Dropout layer to enhance generalization. L2 regularization with a penalty coefficient of 0.0005 is applied to all hidden layers to reduce overfitting. Dropout rates of 40% are used in the first two blocks and 30% in the third. The final layer is a dense layer with 10 units and a softmax activation function, producing probability distributions for multi-class genre classification. Note that each dense layer transforms the

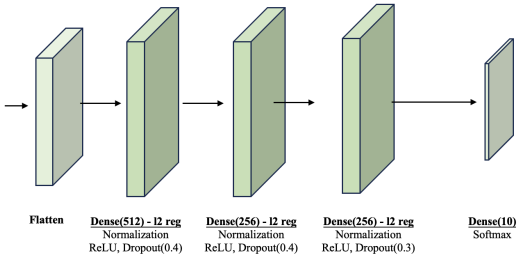

Fig. 3. DNN architecture

previous layer's output using Eq. 1.

$$\text{output} = \text{activation}(W \cdot \text{input} + b) \tag{1}$$

After defining the architecture, the model was compiled using the Adam optimizer with a learning rate of 0.0001. The loss function chosen was sparse categorical cross-entropy, which is appropriate for multi-class classification tasks where the target labels are integers. Model performance was evaluated using accuracy as the primary metric. Training was conducted over 250 epochs with a batch size of 64, and validation data was used to monitor generalization performance during training.

Recurrent Neural Network RNN is designed to process sequential data by retaining information over time through recurrent connections [12]. As music is inherently sequential thus it evolves over time, RNNs (and especially LSTMs) are designed to retain and process temporal information, which helps the model learn patterns that depend on the order and timing of musical elements like rhythm. The implemented RNN model, shown in Figure, begins with two stacked LSTM (Long Short-Term Memory) layers. The first LSTM layer receives the input with shape (130, 13) and is configured to return sequences so that the subsequent LSTM layer can continue processing the temporal dynamics. Following the recurrent layers, the architecture includes a fully connected dense layer with 64 neurons and a ReLU activation function. The final classification layer is a dense layer with 10 units and a softmax activation function, which outputs a probability distribution across the 10 target genres. Unlike the dense network, this architecture inherently preserves temporal structure without the need for flattening the input. Similarly, after defining the architecture, the model was compiled using the Adam optimizer with a learning rate of 0.0001. The loss

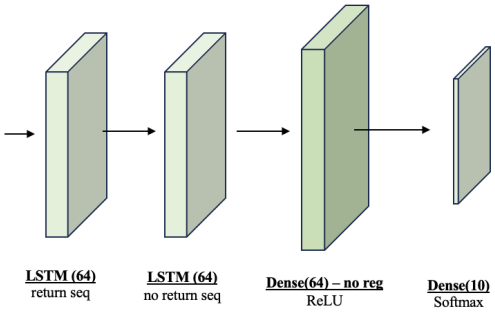

Fig. 4. RNN architecture

function chosen was sparse categorical cross-entropy, which is suitable for multi-class classification tasks where the target labels are integers. Model performance was evaluated using accuracy as the primary metric. Training was conducted over 250 epochs with a batch size of 64, and validation data was used to monitor generalization performance during training.

Convolutional Neural Network CNN consists of a combination of convolution layers and dense layers [5]. Two models were developed and were trained. Unlike in the previous two models, the input data was reshaped first to include a channel dimension to adapt the data to the convolutional layers. The original training and testing sets, used in the previous two models, were modified by adding an additional column before reshaping the input data into the required format.

The base model follows a straightforward architecture composed of three convolutional blocks. As illustrated in Figure 5, each block consists of a Conv2D layer with ReLU activation, followed by a MaxPooling2D layer with stride (2,2) and same padding to reduce spatial dimensions while preserving essential features. The model begins with 32 filters in the first convolutional layer, and increases to 64 filters in the subsequent two layers. After the convolutional layers, the output is flattened and passed through a fully connected dense layer with 64 units before the final softmax output layer, which predicts across 10 classes. This model is compiled using the Adam optimizer with a learning rate of 0.0001 and trained over 250 epochs.

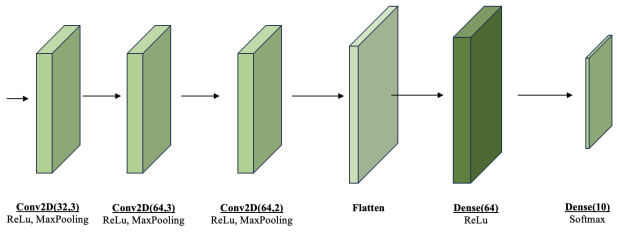

Fig. 5. CNN (base) architecture

The second CNN model, which is the enhanced version shown in Figure 6, builds on the base architecture by incorporating several improvements aimed at increasing generalization and stability during training. Batch normalization is applied after each convolutional layer to standardize intermediate outputs, accelerating training and reducing sensitivity to initialization. Additionally, dropout layers are introduced at various points in the network to mitigate overfitting: 0.2 after the first block, and 0.1 after the second and third convolutional blocks. A final dropout layer of 0.5 is applied after a larger dense layer with 128 units. The model also includes an early stopping callback to halt training if validation loss fails to improve over 20 epochs, helping to prevent overfitting. Like the base model, it is compiled with the Adam optimizer and trained using the same set of parameters.

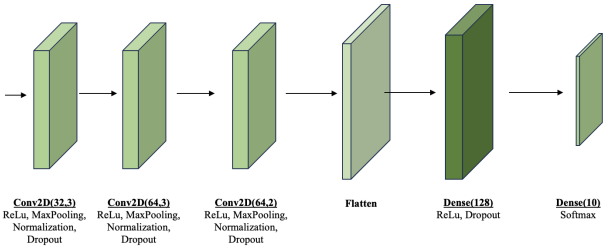

Fig. 6. CNN (enhanced) architecture

IV. Results and Discussion

Table I summarizes the accuracy of these deep learning models applied to genre classification. A dense neural network (DNN) model was tested as it is one of the simplest and fastest neural network architectures to train. However, when dealing with data such as short audio clips, DNNs often fall short in performance compared to models specifically designed to handle spatial patterns [7]. This was evident when DNN achieved the lowest accuracy at 57.0%. Although it can learn some underlying structure in the data, its fully connected layers lack the ability to effectively capture localized features, making it suboptimal for this type of task.

RNN, on the other hand, performed relatively better, achieving an accuracy of 60.9%. RNNs are designed to capture sequential dependencies, making them a natural fit for audio data, which has an inherent temporal structure. However, because the input clips were limited to only three seconds, the ability of the RNN to fully leverage long-term temporal patterns was restricted [8]. Additionally, despite the application of regularization techniques, RNNs are generally more prone to issues like vanishing gradients, which can hinder learning in deeper or more complex architectures. As a result, while the RNN outperformed the DNN, it still did not match the effectiveness of convolutional approaches for this specific task.

CNNs were also trained, as they are better suited for audio-based tasks. Since our preprocessing transformed audio into spectrogram-like visual representations, CNNs became the most appropriate choice. We started with a baseline CNN model and iteratively improved it by incorporating techniques

such as dropout regularization, which prevents overfitting by randomly deactivating nodes during training. This improvement led to the model that performed the best in our experiments, achieving 77. 2% precision. The use of data augmentation techniques, such as time-reversing the audio spectrograms, helped to further diversify the input and boosted model robustness. These strategies contributed significantly to the superior performance of the regularized CNN.

TABLE I
ACCURACY OF DEEP LEARNING MODELS

Model	Accuracy
DNN	0.570
RNN (With Regularization)	0.609
CNN (Without Regularization)	0.670
CNN (With Regularization)	0.772

The final CNN model demonstrated strong capability in extracting and learning the complex audio features necessary for genre classification. Its performance highlights the importance of transforming raw audio into formats that align with a CNN's strength which is image-like representations. This conversion allows the model to interpret audio as structured visual patterns, enabling it to detect subtle genre-specific traits and improving its generalization to unseen audio files.

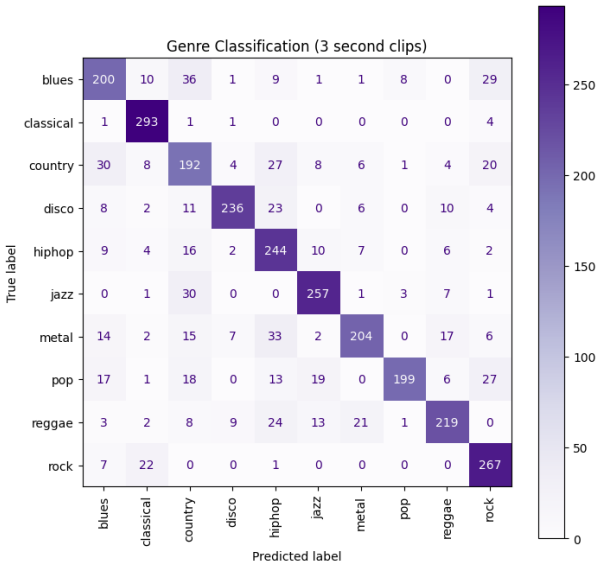

Fig. 7. CNN (Enhaced) Confusion Matrix

The confusion matrix in Figure 7, generated by the best model (enhanced CNN model) using test data, offers a deeper understanding of the behavior of the classifier in specific genres. It shows strong performance for clearly distinguishable genres like classical, jazz, and rock, which had high true positive rates (e.g., classical with 293, rock with 267). However, genres such as country, metal, and reggae experienced more frequent misclassifications. For example, country was often confused with blues and disco, which may suggest overlapping musical characteristics or insufficient model sensitivity to subtle genre-specific cues. Additionally, some genres like pop and reggae appear to be confused with a broader range of classes, possibly indicating their hybrid nature or more complex acoustic signatures. Nevertheless, this further demonstrates the enhanced CNN model's ability to distinguish between a wide variety of musical styles, validating the approach of using regularized convolutional architectures on short audio clips transformed into spectrogram-like inputs.

V. CONCLUSION

Among the models developed, the enhanced CNN model achieved the highest accuracy at 77.2%. This result highlights the effectiveness of regularization techniques in enhancing model generalization, particularly in convolutional architectures that can learn complex patterns in the audio features that are relevant for music genre classification. In contrast, DNN performed the worst, achieving only 57% accuracy, suggesting that fully connected architectures might not be sufficient to capture local patterns in audio data. The RNN, though benefiting from regularization, still lagged behind CNN, likely due to the limited temporal information provided by the 3-second audio clips.

To further enhance the model's performance, future work could explore deeper architectures or hybrid models that combine convolutional and recurrent layers. Incorporating more diverse data augmentation techniques or using longer audio segments may also improve genre distinction. Ultimately, fine-tuning hyperparameters and experimenting with different regularization strategies could also lead to better generalization across varied characteristics of music genres.

REFERENCES

[1] Adnan, K., & Akbar, R. (2019). An analytical study of information extraction from unstructured and multidimensional big data. Journal of Big Data, 6(1), 1-38.

[2] Beer, D. (2013). Genre, boundary drawing and the classificatory imagination. Cultural Sociology, 7(2), 145-160.

[3] Buisson, M., Alonso-Jiménez, P., & Bogdanov, D. (2022, May). Ambiguity modelling with label distribution learning for music classification. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 611-615). IEEE.

[4] Fu, Z., Lu, G., Ting, K. M., & Zhang, D. (2010). A survey of audio-based music classification and annotation. IEEE transactions on multimedia, 13(2), 303-319.

[5] Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. Artificial intelligence review, 53, 5455-5516.

[6] Li, G., Zhang, M., Li, J., Lv, F., & Tong, G. (2021). Efficient densely connected convolutional neural networks. Pattern Recognition, 109, 107610.

[7] Nazari, F., & Yan, W. (2021). Convolutional versus dense neural networks: Comparing the two neural networks performance in predicting building operational energy use based on the building shape. arXiv preprint arXiv:2108.12929.

[8] Pfalz, A. (2018). Generating Audio Using Recurrent Neural Networks. Louisiana State University and Agricultural & Mechanical College.

[9] Ramírez, J., & Flores, M. J. (2020). Machine learning for music genre: multifaceted review and experimentation with audioset. Journal of Intelligent Information Systems, 55(3), 469-499.

[10] Sanden, C., & Zhang, J. Z. (2011, July). Enhancing multi-label music genre classification through ensemble techniques. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (pp. 705-714).

[11] Trohidis, K., Tsoumakas, G., Kalliris, G., & Vlahavas, I. (2011). Multi-label classification of music by emotion. EURASIP Journal on Audio, Speech, and Music Processing, 2011, 1-9.

[12] Zargar, S. (2021). Introduction to sequence learning models: RNN, LSTM, GRU. Department of Mechanical and Aerospace Engineering, North Carolina State University.