Genre-Based Classification of Songs Using Deep Learning Models

Martin Donaire

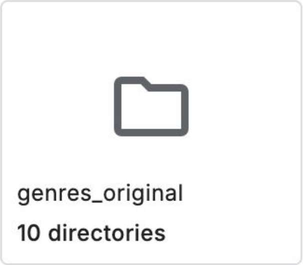
Introduction

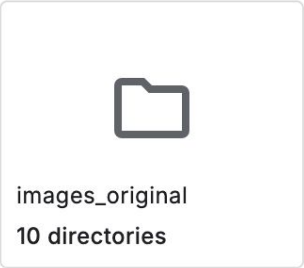
Challenge: Music genres often overlap, making classification complex.

Objective: Classify music genres using deep learning models.

Why It Matters: Enhances music recommendation, playlist curation, and analysis.

Applications: Improves music streaming platforms, user experience, and music industry analytics.





Data Preprocessing

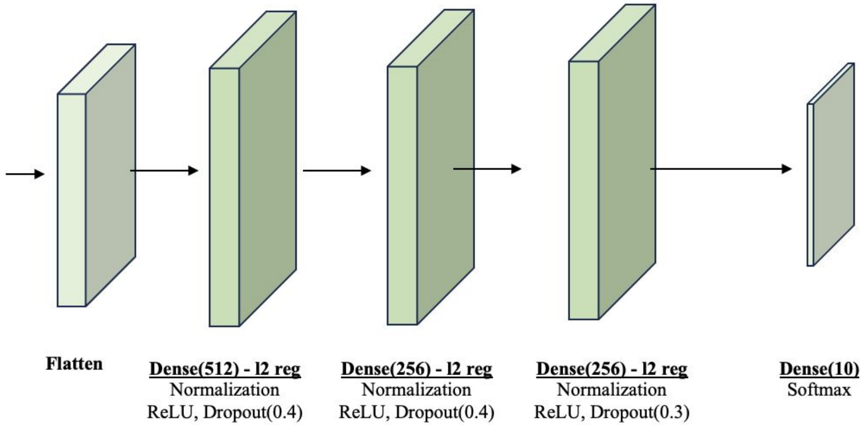
Dataset: GTZAN from Kaggle, 10 genres (e.g., hip-hop, rock), 100 audio files (30s each) per genre. It was preprocessed as follows:

- **Segmentation**: Split each 30s track into ten 3s segments.
- **MFCC Extraction**: Transform segments into Mel Frequency Cepstral Coefficients (MFCCs) to capture timbral/spectral features.
- **MFCC Settings**: 22,500 Hz sampling rate, 2048 FFT window, 512 hop length, 13 MFCCs per frame.
- **Purpose of MFCCs**: Compresses audio into perceptually relevant, image-like features for deep learning.

Data Splitting: 70% training (30% validation), 30% testing, *stratified*.

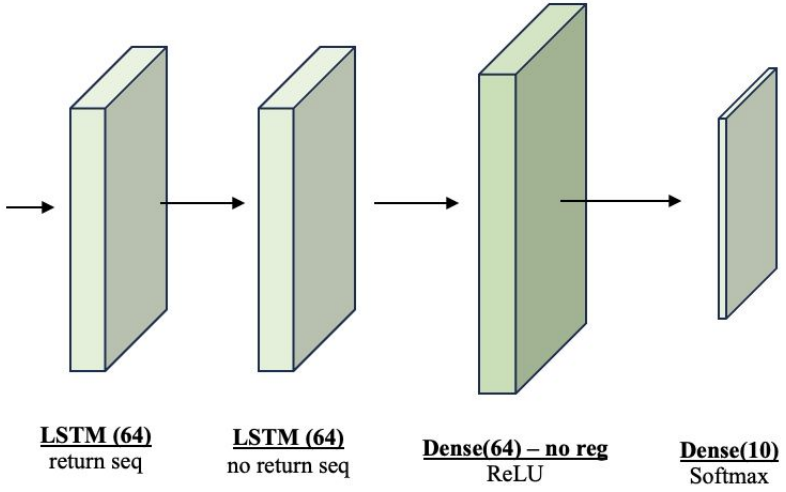
Model 1: Dense Neural Network

- **Architecture**: Fully connected layers (512, 256, 64 neurons), flattens MFCC input (130x13) into 1D array.
- **Features**: ReLU activation, L2 regularization (0.0005 penalty), dropout (40%, 30%), softmax for 10 genres.
- **Purpose**: Simple and fast but struggles with localized audio patterns.
- **Training**: Adam optimizer (0.0001 learning rate), sparse categorical cross-entropy, 250 epochs, batch size 64.



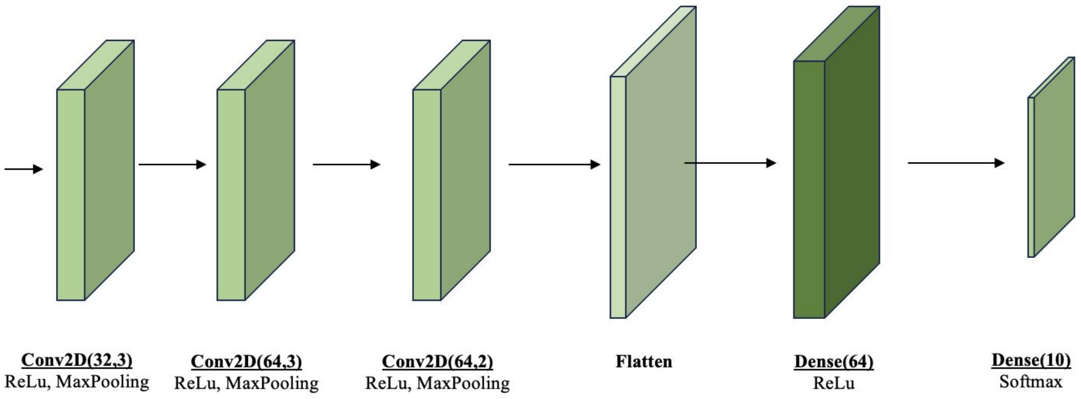
Model 2: Recurrent Neural Network

- **Architecture**: Two stacked LSTM layers for sequential MFCC input (130x13), 64-unit dense layer, softmax output.
- **Features**: Captures temporal dependencies, uses ReLU activation, softmax for 10 genres.
- **Purpose**: Ideal for sequential audio data, limited by short 3s clips.
- **Training**: Adam optimizer (0.0001 learning rate), sparse categorical cross-entropy, 250 epochs, batch size 64.



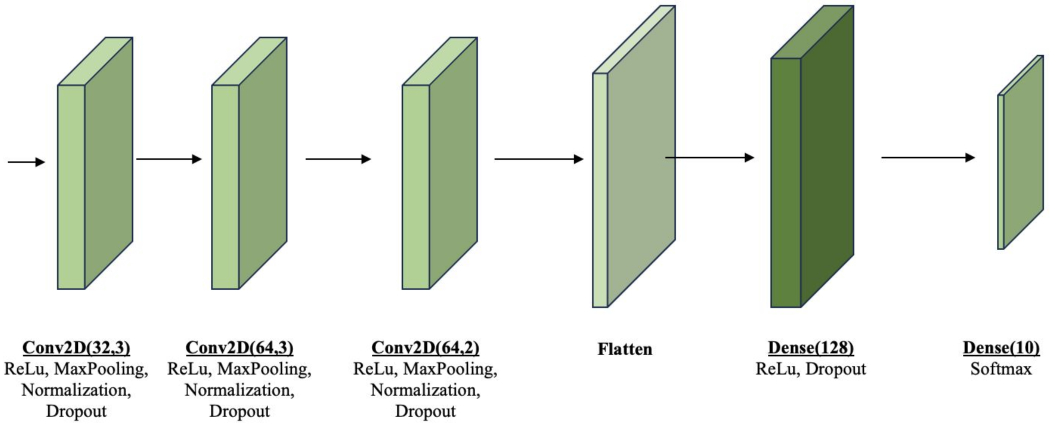
Model 3: Convolutional Neural Network (Base)

- **Architecture**: Three convolutional blocks (32, 64 filters), ReLU activation, max-pooling (stride 2x2), same padding.
- **Structure**: Flattens output, 64-unit dense layer, softmax for 10 genres.
- **Purpose**: Captures spatial patterns in MFCCs, treating them as images.
- **Training**: Adam optimizer (0.0001 learning rate), sparse categorical cross-entropy, 250 epochs, batch size 64.



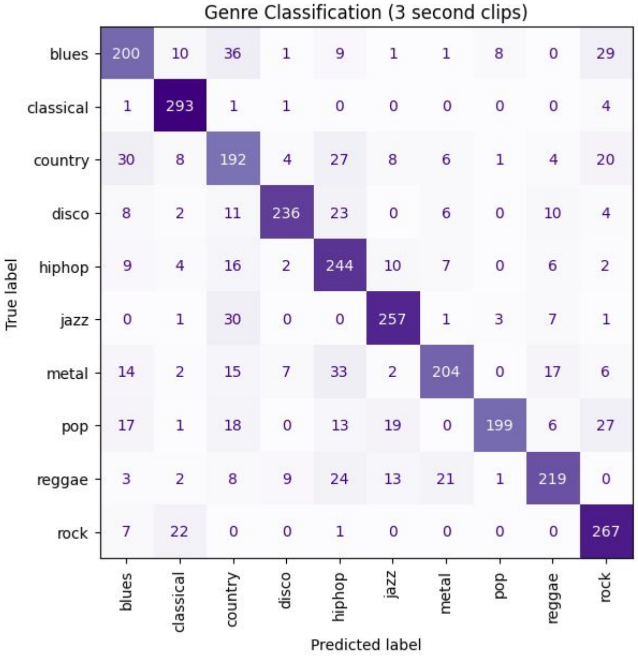
Model 4: Convolutional Neural Network (Enhanced)

- **Architecture**: Builds on Base CNN with three convolutional blocks (32, 64 filters), ReLU, max-pooling.
- **Enhancements**: Batch normalization after each conv layer, dropout (0.2, 0.1, 0.5), 128-unit dense layer, early stopping (20 epochs).
- **Purpose**: Improves generalization and stability, excels at complex audio patterns.
- **Training**: Adam optimizer (0.0001 learning rate), sparse categorical cross-entropy, 250 epochs, batch size 64.



Model Performance

- **Accuracy Results**:
  - DNN: 57.0% (weakest, poor at localized patterns).
  - RNN: 60.9% (better, limited by 3s clips, vanishing gradients).
  - Base CNN: 69.9% (effective for spatial patterns).
  - Enhanced CNN: 77.2% (best, due to regularization, data augmentation).
- **Key Insight**: Enhanced CNN excels by treating MFCCs as images, boosted by dropout and time-reversed spectrograms.
- **Confusion Matrix**: Strong for classical (293 TP), rock (267 TP); misclassifications in country (with blues), pop (multiple genres).



Summary and Future Direction

- **Key Finding**: *Enhanced CNN achieved 77.2% accuracy,* outperforming DNN (57.0%), RNN (60.9%), and Base CNN (69.9%).
- **Implication**: Regularized CNNs are highly effective for genre classification using MFCCs.
- **Limitations**: Short 3s clips limit RNNs; genre overlap causes misclassifications.
- **Future Work**:
  - Explore deeper or hybrid models (e.g., CNN+RNN).
  - Use longer audio segments or advanced data augmentation.
  - Fine-tune hyperparameters for better generalization.
- **Takeaway**: MFCCs with regularized CNNs offer a robust approach for music genre classification.