ML-Based Audio Classification: Musical Genre Classification

Prepared by: Liul Teshome

Music genre classification—the process of automatically classifying a piece of music. Is a vital for music streaming and content organization. This project uses deep learning to classify 10 genres from the GTZAN dataset, employing Mel spectrograms, a custom CNN, and a modified ResNet18. The custom CNN achieved 89.49% test accuracy, while ResNet18 reached 96.80%. This report details the methodology, results, and insights, providing a robust framework for genre classification.

Dataset

The GTZAN dataset comprises 1,000 WAV files, each ~30 seconds, sampled at 22,050 Hz with 16-bit PCM, across 10 genres, with 100 files per genre. Its balance and diversity make it ideal for classification tasks.

Data Preprocessing

To ensure data quality and consistency, the following preprocessing steps were performed:

- **Exploratory Analysis**: Audio files were audited to gain insights into their characteristics. Waveforms and Mel spectrograms were visualized to understand temporal and frequency patterns.
- **File Format Verification**: All files were confirmed to be in WAV format.
- **Sampling Rate Check**: The sampling rate was consistent at 22,050 Hz
- **Bit Depth Check**: All files used 16-bit PCM encoding.
- **Duration Consistency**: Audio durations varied slightly (Min=29.93s, Max=30.65s, Avg=30.02s). To address this, files were padded with zeros or trimmed to a uniform 30 seconds.

Feature Extraction: Mel spectrograms were extracted due to their effectiveness for CNNs:

- **Segmentation**: Split 30-second files into 3-second segments with 1.5-second overlap, yielding ~19,980 segments to capture local patterns and ensure no transitions were missed.
- **Mel Spectrograms**: Computed using STFT, Mel scale mapping, with parameters: 128 Mel bands, hop length 512, FFT size 2,048. Output shape: (num_segments, 128, 130, 1).
- **Why Mel Spectrograms?**: Their 2D, perceptually relevant format suits CNNs, outperforming MFCCs or chroma. And Research Support Mel Spectrograms.
- **Split**: Stratified into training (15,984 samples, 80%), validation (1,998, 10%), and test (1,998, 10%) sets.

Model Development: A custom CNN was developed from scratch to classify the 10 genres, designed to balance simplicity and performance.

Architecture: The custom CNN processes Mel spectrograms (1, 128, 130) through three convolutional layers: Conv1 (32 filters, 3x3, ReLU, MaxPool2d) outputs (32, 64, 65); Conv2 (64 filters) outputs (64, 32, 32); Conv3 (128 filters) outputs (128, 16, 16). Global average pooling yields (128,), followed by 0.5 dropout and a linear layer (128, 10) for 10 genres.

Why This Design? Simple yet effective, with global average pooling to reduce overfitting and dropout for generalization.
Comparison: Unlike traditional ML (limited by handcrafted features), RNNs (heavy for spectrograms), or transformers (data-intensive), CNNs excel at 2D spectrogram processing.

Training

- **Setup**: PyTorch, batch size 64, Adam optimizer (lr=0.001), Cross-Entropy Loss, 50 epochs.
- **Process**: Trained on 15,984 samples, validated on 1,998, with loss and accuracy tracked to monitor convergence.

Custom CNN Model Result: Evaluated using accuracy, precision, recall, and F1-score, suitable for the dataset:

Metric	Validation (%)	Test (%)
Accuracy	89.94	89.49
Precision	90.33	89.97
Recall	89.94	89.50
F1-Score	89.94	89.56

Transfer Learning: To explore the potential of pre-trained models, a modified ResNet18 was trained and evaluated on the same dataset, as research suggests it excels in audio classification tasks. Modification with Conv2d(1, 64, 7, stride=2, padding=3), output Linear(512, 10)

Metric	Validation (%)	Test (%)
Accuracy	96.30	96.80
Precision	96.39	96.87
Recall	96.30	96.80
F1-Score	96.29	96.80

Results

The custom CNN achieved a test accuracy of 89.49%, while the modified ResNet18 reached 96.80% on the GTZAN dataset for 10-genre music classification. In comparison, human performance, as studied by Perrot and Gjerdigen (1999), showed college students correctly identifying genres at 53% accuracy with 250-ms audio samples and 70% with 3-second samples in a 10-way forced-choice task (chance level: 10%). Both models significantly outperform human accuracy at 3 seconds, with the custom CNN surpassing humans by ~19% and ResNet18 by ~27%. From the two model, ResNet18 offering near-state-of-the-art performance.

Observations: Mel spectrograms with segmentation were key to success, increasing data and capturing local patterns.

Future work: Explore hybrid features (e.g., MFCCs) and develop data augmentation (e.g., pitch shifting). Create recurrent neural networks (RNNs) and Transformers Model to capture temporal dependencies and long-range audio patterns for improved accuracy.