Hard Path Pipeline

Hard Path: End-to-End Multi-Input Deep Learning with Vocal & Audio + Sequential Models

Pipeline Outline

1. Data Input & Preprocessing:
- Load raw audio for vocal and non-vocal (instrumental/audio) tracks separately.
- Segment into fixed-length frames if necessary.
- Extract time-frequency features (Mel-Spectrogram or raw waveform) for both vocal and audio tracks.
- Normalize features independently.

2. Metadata Preparation:
- Encode metadata (Artist, Tempo, Genre) as embeddings or one-hot vectors.

3. Model Architecture:
- Dual-input deep model:
- Branch 1: CNN or Transformer layers on vocal track features.
- Branch 2: CNN or Transformer layers on audio track features.
- Branch 3: Metadata input processed via dense layers.
- Concatenate outputs of three branches.
- Fully connected layers for final classification.

4. Training Setup:
- Use advanced augmentation on vocal/audio tracks separately.
- Use Nested Cross-Validation or Bayesian Optimization for hyperparameter tuning.
- Apply dropout, batch normalization, and early stopping.

5. Evaluation:
- Calculate accuracy, macro-averaged F1-score, confusion matrix.
- Conduct ablation to evaluate vocal-only, audio-only, and combined model contributions.
- Perform error analysis per language and per input type.

Handling Vocal/Audio Tracks:
- Use separate model branches for vocal and audio inputs.
- Normalize and augment vocal and audio inputs independently.
- Integrate metadata as auxiliary input.

Machine Learning Techniques Summary

ML Techniques Used:
- Deep Learning Architectures (CNN, Transformer, RNN)
- Nested Cross-Validation
- Bayesian Optimization
- Dropout, Batch Normalization, Early Stopping
- Accuracy, Macro-averaged F1-score, Confusion Matrix
- Ablation Studies
- Embeddings for metadata

- Data Augmentation on multi-input branches