Medium Path Pipeline

Medium Path: Time-Frequency Features on Vocal & Audio Tracks + CNN

Pipeline Outline

1. Data Input & Preprocessing:
- Load separate vocal and audio waveforms for each sample.
- Extract Mel-Spectrograms or STFT for both tracks individually.
- Normalize or standardize spectral features per track.

2. Feature Integration:
- Stack vocal and audio spectrograms as separate input channels.
- Or concatenate flattened vocal and audio feature vectors before feeding into model.

3. Data Augmentation:
- Apply augmentations such as time-shifting, pitch shifting, or noise addition on vocal track spectrograms.
- Optionally augment audio track spectrograms separately.

4. Modeling:
- Build CNN architecture that:
- Takes vocal track spectrogram as input.
- Takes audio track spectrogram as a second input channel or branch.
- Fuse learned features before final classification layer.

5. Training & Validation:
- Use K-Fold Cross-Validation or Train-Test split.
- Employ early stopping based on validation loss.

6. Evaluation:
- Report accuracy, macro F1.
- Perform ablation to compare vocal-only vs combined inputs.

Handling Vocal/Audio Tracks:
- Extract and normalize vocal and audio spectrograms separately.
- Feed as separate channels or inputs into CNN.

Machine Learning Techniques Summary

ML Techniques Used:
- Convolutional Neural Networks (CNN)
- K-Fold Cross-Validation
- Early Stopping
- Accuracy, Macro F1-score
- Ablation Studies
- Feature Normalization (StandardScaler, RobustScaler)
- Data Augmentation (time-shifting, pitch-shifting, noise addition)