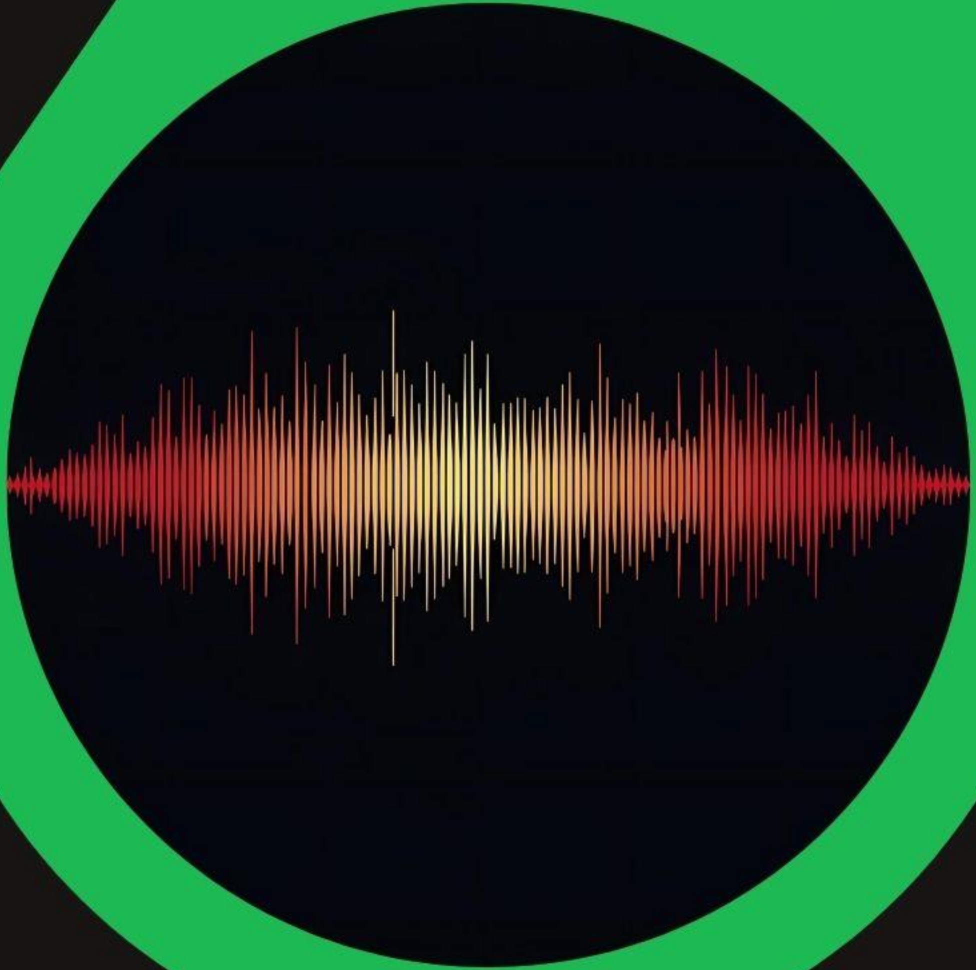
TEAM AESTHEVAL

AesthEval: Predicting song quality through audio analysis

Team AesthEval: Aryan Daivik Lakshmi Shourya



Problem Statement

The Challenge

How do you predict if a song sounds "good" just by analyzing the audio itself? No reading lyrics, no checking who the artist is, no user data like streaming numbers just the audio features.

Why It Matters

Small Independent Creators and AI composers need automatic quality assessment to evaluate their output without need of expensive pro feedback

Real-World Impact

Spotify gets over 100,000 new songs uploaded every single day. Having human experts listen to each one is impossible but this is needed to help new artists understand what's working, and assist music labels in discovering talent.

Scientific Goal

Decoding which measurable sound properties drive quality perception. Is it smoothness of the vocals, the rhythmic precision, or the melodic structure ?

Objectives & Scope

We want to figure out which parts of a song matter most for quality. and to develop a robust framework for multi-dimensional assessment.

01

Complete Processing Pipeline

Develop an audio processing pipeline by taking any song, breaking it down into its building blocks (vocals, instruments, drums), and extracting meaningful measurements from each part.

02

Test Different Ways to Merge Information

We're testing three approaches: Early Fusion (mix everything upfront), Late Fusion (analyze separately, then vote), and Hybrid (a smart middle ground).

03

Find What Matters Most

Identify which audio features: vocal characteristics, harmonic content, or rhythmic patterns are most predictive of perceived quality across five aesthetic dimensions.

Related Work

Existing Approaches

Most music recommendation systems work like this: "People who liked Song A also liked Song B, so you'll probably like it too." That's collaborative filtering.

While these approaches work well for personalized recommendations, they don't explain *why* certain songs are perceived as high quality from an acoustic perspective, basically we need something objective also but something that also works

Our Contribution

There's surprisingly little research that breaks songs into their musical components (vocals, harmony, rhythm) and asks: "Which elements contribute most to perceived quality?" Most studies either analyze whole songs or focus on technical tasks like genre classification rather than quality prediction, or they use user data like streaming number

Some existing work for music evaluation

For Collaborative Filtering & Music Recommendation:

- Schedl et al. (2018) - "Current Challenges and Visions in Music Recommender Systems Research"

For Hit Song Prediction:

- Pachet & Roy (2008) - "Hit Song Science Is Not Yet a Science"

Methodology: Dataset

Our project is built on the **SongEval Dataset**, a comprehensive data collection specifically designed for music perception quality assessment.

Dataset Scope

2,399 complete songs spanning approximately 140 hours of audio content across 9 distinct musical genres.

Expert Annotations

16 expert annotators with musical training have rated each song on a 5-point scale, ensuring reliable ground-truth labels.

Five Aesthetic Dimensions

Coherence

Structural unity and flow

Musicality

Melody & harmony quality

Memorability

Catchiness & recall

Clarity

Section distinctness

Naturalness

Vocal realism

Methodology: Processing Pipeline

We employ a systematic 4-step pipeline to transform raw audio into aesthetic quality predictions.

Source Separation

Decompose the audio into vocals, accompaniment, harmonic, and percussive tracks for focused analysis using libraries like Spleeter and librosa.

Feature Extraction

Extract key acoustic features like MFCCs, Chroma, Spectral Centroid, Zero Crossing Rate from each separated track using specialized libraries.

Multimodal Fusion

Integrate feature vectors from different sources using early, late, or hybrid fusion, maximizing information from all modalities.

Quality Prediction

Train regression models to predict the 5 aesthetic ratings using scikit-learn ensemble methods.

Work Plan – Train – Test Split

Shourya Goyal

Data Pipeline & Feature Extraction

- Build data loading infrastructure because of the big size
- Implement feature extraction

Aryan Gupta

Source Separation & Model Training

- Configure Spleeter for stem separation
- Train regression models – Polynomial, KNN-based, RF
- Optimize hyperparameters

Lakshmi C

Fusion Strategies & Analysis

- Implement fusion techniques
- Analyze modality importance
- Compare strategy performance

Daivik Gupta

Evaluation & Documentation

- Design evaluation protocols
- Document findings and results
- Per-dimension Analysis

Deliverables & Evaluation

Project Deliverables

01

Runnable Python Pipeline

A complete, documented codebase capable of processing any audio file and extracting all multi-modal features.

02

Modality Analysis Report

Comprehensive analysis identifying which features are most predictive (e.g., "Vocal features strongly predict Naturalness scores").

03

Fusion Strategy Comparison

Quantitative comparison showing performance differences between fusion approaches.

Evaluation Methodology

Primary Metric

Root Mean Square Error (RMSE) will measure prediction accuracy against human expert scores for each aesthetic dimension.

Validation Approaches

- **Cross-validation:** 5-fold cross-validation to ensure robust performance estimates.
- **Per-dimension Analysis:** Evaluate performance separately for each aesthetic dimension to identify strengths and weaknesses.

Challenges & Mitigation Strategies

We have identified three primary technical challenges and developed mitigation approaches for each.

140+ HOURS AUDIO	POOR FEATURES	OVERFITTING
Computational Cost	Separation Quality	High Dimensionality

Challenges & Mitigation Strategies

We have identified three primary technical challenges and developed mitigation approaches for each.

Computational Cost

Challenge:
- 140 hours of audio: High computational resources & time.
- Risk: Development delays.

Mitigation:
- Parallel Processing (multi-core).
- Cache extracted features (disk reuse).
- Prioritize efficient algorithms.

Separation Quality

Challenge:
- Source Separation Imperfections
- Risk: Artifacts (bleeding, phase distortion) may affect feature quality.

Mitigation:
- Select features known to be robust to separation artifacts.
- Validate separation quality on a subset before full-scale processing.

High Dimensionality

Challenge:
- Large feature vectors (many features, multiple stems).
- Risks: Overfitting.

Mitigation:
- Feature Selection techniques.
- Dimensionality Reduction (e.g., PCA).
- Apply Regularization.