# Baseline Pipeline Implementation for Automatic Song Aesthetics Evaluation

Ahsan Adil, Muhammad Hamza Malik, Abdullah Azmat
*SEECS*
*National University of Science and Technology*
Islamabad, Pakistan
{aadil, mmalik, aazmat}.bscs24seecs@seecs.edu.pk

***Abstract*—This report details the implementation and comparison of three baseline models for Assignment 2 of the CS-272 semester project, aligned with the ICASSP 2026 Automatic Song Aesthetics Evaluation Challenge (Track 2). The objective is to predict five dimensions of musical aesthetics from raw audio. We successfully implemented a robust data processing pipeline capable of handling large audio files by streaming data and processing random 10-second chunks. We compare three distinct baselines: (1) a "classic" Random Forest using librosa features, (2) a Transformer model pre-trained on speech (Wav2Vec2), and (3) a Transformer model pre-trained on general audio (AST). Our results clearly demonstrate the superiority of the AST model, which achieved an average Spearman's Correlation (SRCC) of 0.828, significantly outperforming the Random Forest (0.540) and the speech-based model (0.100). This indicates that the pre-training domain is a critical factor for success in this task.**

**GitHub Repository:** https://github.com/SkinnyLadd/ASLP-SongEval

## I. Introduction

Evaluating the aesthetic quality of music is a complex, subjective task, yet it is crucial for applications in music generation and recommendation. The ICASSP 2026 Grand Challenge on Automatic Song Aesthetics Evaluation (Track 2) provides a benchmark for this task, asking participants to develop models that can predict human-perceived aesthetic scores.

The task is a multi-output regression problem: given a full-length song, our model must predict five continuous scores: Coherence, Memorability, Naturalness, Clarity, and Musicality.

This report fulfills the requirements for Assignment 2, "Baseline Pipeline Implementation." We describe our data pipeline and present a comparative analysis of three baseline models. The goal is to establish a strong foundation and identify the most promising architecture for our "Proposed Solution" in Assignment 3.

## II. Dataset and Methodology

### *A. Dataset*

We use the `ASLP-lab/SongEval` dataset provided for the challenge. This dataset contains over 2,400 full-length songs. Crucially, each song is annotated by multiple professional annotators, providing a list of scores for each of our five target dimensions.

### *B. Data Processing Pipeline*

A significant technical challenge was the memory constraints of the development environment (Google Colab). Loading and processing full-length (2-5 minute) audio files for a large dataset leads to repeated std::bad_alloc (Out of Memory) crashes.

To overcome this, we designed a highly memory-efficient pipeline with the following steps:

1) **Stream Loading:** We load the dataset from Google Drive using `streaming=True` to avoid loading the entire dataset manifest into RAM.
2) **Deferred Decoding:** We cast the audio column to `datasets.Audio(decode=False)`. This prevents the library from automatically loading full audio files into memory.
3) **Manual Chunking:** We iterate through the dataset manually. For each song, we use `librosa.load()` with `offset` and `duration` parameters to load only one random 10-second chunk. This keeps RAM usage minimal and stable.
4) **Label Generation:** For each example, we parse the list of `annotation` dictionaries and compute the mean score for each of the five aesthetic dimensions. This mean value serves as our ground truth label.

This pipeline allows us to process a dataset of any size within the 12GB RAM limit of our environment.

### *C. Baseline Models*

We implemented three baselines to provide a comprehensive comparison.

*1) Baseline 1: Random Forest (Classic ML):* This baseline tests the efficacy of traditional, hand-crafted audio features. We use `librosa` to extract MFCCs (20), Chroma features (12), and Spectral Contrast (7) from the 10-second audio chunk. These features are averaged across the time axis, concatenated into a 49-dimensional vector, and used to train a `scikit-learn` `RandomForestRegressor` (100 estimators).

*2) Baseline 2: Wav2Vec2 (Speech-Domain DL):* This baseline tests a large Transformer model pre-trained exclusively on speech data (`facebook/wav2Vec2-base`). We hypothesize that features learned from speech will not transfer well to music. We replace the model's classification head with a 5-output regression head and fine-tune it on our 10-second chunks.

*3) Baseline 3: AST (Audio-Domain DL):* This baseline uses the Audio Spectrogram Transformer (`MIT/ast-finetuned-audioset-10-10-0.4593`), which was pre-trained on AudioSet, a large dataset of general sounds that includes music. We hypothesize this model's features will be far more relevant. We replace its 527-output classification head with a 5-output regression head for our task.

## III. Experiments and Results

### A. Experimental Setup

All models were trained and evaluated on data processed by our pipeline. For the deep learning models, we used a custom `RegressionTrainer` with a Mean Squared Error (MSE) loss function. All training was conducted on a Google Colab T4 GPU.

We use two primary metrics for evaluation:

- **Mean Absolute Error (MAE):** The average absolute difference between our predictions and the mean annotator scores. Lower is better.
- **Spearman's Correlation (SRCC):** Measures the monotonic relationship between our predicted rankings and the ground truth. This is our primary metric for success. A score of +1.0 is a perfect correlation, while 0.0 is random.

### B. Results

The final evaluation results for our three baselines are presented in Table I. Due to different processing requirements and stability, dataset sizes for this initial comparison were varied.

TABLE I
Comparison of Baseline Model Performance

| Model | n | Train Time | MAE ↓ | Avg. SRCC ↑ |
|---|---|---|---|---|
| B1: Random Forest | 500 | ~3 min | 0.656 | 0.540 |
| B2: Wav2Vec2-base | 300 | ~6 min | 0.936 | 0.100 |
| B3: AST | 500 | ~7 min | **0.505** | **0.828** |

## IV. Discussion and Analysis

The results in Table I provide a clear and compelling story.

**Baseline 1 (Random Forest)** performed surprisingly well, achieving an SRCC of 0.540. This confirms that traditional audio features like MFCCs and chroma do capture meaningful information related to musical aesthetics.

**Baseline 2 (Wav2Vec2)** was a near-complete failure. With an SRCC of 0.100, its predictions are barely better than random. The MAE of 0.936 is also the highest. This strongly supports our hypothesis: features learned from speech do not generalize to the abstract, harmonic, and rhythmic complexities of musical aesthetics.

**Baseline 3 (AST)** was the undeniable winner. It achieved the lowest MAE (0.505) and a very strong average SRCC of 0.828. This indicates a high degree of correlation with human judgments. Analyzing its per-dimension scores reveals this strength is consistent:

- **Coherence:** 0.855 SRCC
- **Memorability:** 0.848 SRCC
- **Naturalness:** 0.780 SRCC
- **Clarity:** 0.852 SRCC
- **Musicality:** 0.804 SRCC

This success is almost certainly due to AST's pre-training on AudioSet, which provided it with a rich feature representation for general audio, including music.

## V. Conclusion

For Assignment 2, we successfully implemented a robust, memory-safe data processing pipeline and compared three distinct baseline models. The results conclusively show that the choice of pre-training domain is the most critical factor. The speech-based Wav2Vec2 model failed, while the general-audio-based AST model (Baseline 3) proved highly effective, achieving a strong correlation (0.828 SRCC) with human aesthetic ratings.

For Assignment 3, the AST model will serve as our strong baseline, which we will aim to improve through more advanced data chunking strategies (e.g., averaging predictions over the entire song) and hyperparameter tuning.

## VI. Author Contributions

As required, the contribution of each team member for this assignment is detailed in Table II.

TABLE II
Author Contribution Table

| Member | Tasks Completed |
|---|---|
| Ahsan Adil | Pipeline Debugging, Baseline 3 (AST) Implementation |
| Abdullah Azmat | Baseline 1 (RF) Implementation, Report Writing |
| Muhammad Hamza Malik | Baseline 2 (Wav2Vec2) Implementation, Results Analysis |

## References

[1] Y. A. B. C. Yao, et al. "SongEval: A Benchmark Dataset for Song Aesthetics Evaluation." *arXiv preprint arXiv:2505.10793*, 2025. (Placeholder for the challenge paper)

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. "wav2vec 2.0: A framework for self-supervised learning of speech representations." *Advances in Neural Information Processing Systems*, 33, 2020.

[3] S. Gong, Y. Lo, and J. Glass. "AST: Audio Spectrogram Transformer." *Interspeech*, 2021.