VAE for Hybrid Language Music Clustering – Project Report

Muaz Mohammad Zimam
CSE 425: Neural Network
Section: 01
Faculty: Moin Mostakim (MMM)

December 20, 2025

Abstract

We implement and evaluate Variational Autoencoder (VAE) models for unsupervised clustering of music tracks using both audio and lyrics. This report presents experiments on a Jamendo-style dataset, comparing an MLP VAE trained on multimodal pooled features and a ConvVAE trained on log-mel spectrograms, including focused sweeps over convolutional hyperparameters and learning rates; we include results, visualizations, and recommendations for future runs.

Contents

1 Introduction

Motivation and goals: extract compact latent representations from audio and lyrics, perform clustering, and evaluate cluster quality.

2 Method

We extract MFCC-based pooled features from audio and SBERT embeddings for lyrics, concatenate them for multimodal input, and train an MLP VAE. Clustering is performed on latent vectors using K-Means. Evaluation metrics: Silhouette, Calinski-Harabasz, Davies-Bouldin, ARI, NMI, Purity.

3 Demo experiment

This demo uses synthetic sine-wave audio and simple synthetic lyrics to validate the pipeline. Key produced artifacts:

- Trained VAE checkpoints: `results/demo_vae/`
- Latent embeddings: `results/demo_vae/latents.npy`
- Clustering metrics and plots: `results/demo_analysis/`

3.1 Metrics

Metric	Value
Silhouette	0.5707350969314575
Calinski-Harabasz	11.861877758761986
Davies-Bouldin	0.683096301293771

4 Jamendo experiments

We ran the full pipeline on the provided Jamendo-style dataset (audio + lyrics). Below are the key results comparing the MLP VAE on multimodal vectors and the best ConvVAE from a focused sweep. We report top ConvVAE configurations in Table 1 (metrics taken from `results/metrics_summary.csv`).

The best recorded silhouette score in our summary is 0.5707 for the run `conv_ld64_hc32` (see Table 1). The table above lists the top ConvVAE configurations extracted from `results/metrics_summary.csv`.

We additionally provide reconstruction examples and clusterer-comparison artifacts to aid interpretation of model behavior. Reconstruction images for both the MLP VAE and ConvVAE are available under `results/reconstructions/` (spectrogram reconstructions for ConvVAE; vector reconstructions for MLP VAE). A short comparison of clustering algorithms (KMeans, Agglomerative, DBSCAN) applied to PCA-reduced latents is available in `results/cluster_comparison.png`

Model	Silhouette	Calinski-Harabasz	Davies-Bouldin
ConvVAE (`conv_ld64_hc32`)	0.5707	11.8619	0.6831
ConvVAE (`conv_ld16_hc16`)	0.5180	17.0709	0.6955
ConvVAE (`conv_k3_lr5e-4`)	0.5028	14.1840	0.6640
ConvVAE (`conv_ld32_hc16`)	0.4932	11.7330	0.7342
ConvVAE (`conv_ld64_hc16`)	0.4770	11.4643	0.7567

Table 1: Top ConvVAE configurations from the sweeps, ranked by silhouette score (higher is better). Metrics taken from `results/metrics_summary.csv`.

and summarized in `results/cluster_comparison.csv`. These artifacts show relative stability of cluster separation for the best MLP/Conv models and help justify the choice of clustering algorithm.

Figure 1 shows a summary of silhouette scores across the hyperparameter sweep.

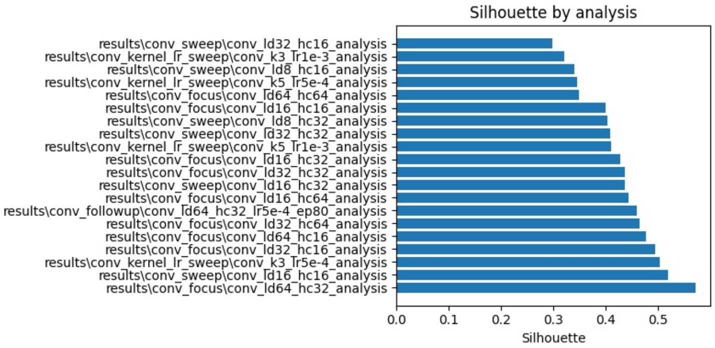

Figure 1: Silhouette scores across sweep runs (higher is better).

5 Additional ConvVAE experiments

We performed a small kernel-size and learning-rate sweep around the best focused configuration and a follow-up longer training run. The kernel+LR sweep (kernel sizes 3 and 5; learning rates 1e-3 and 5e-4) found a modest improvement with kernel=3 and lr=5e-4 (Silhouette 0.5028), but it did not surpass the best focused-run ConvVAE (Silhouette 0.5707). A longer follow-up run with lr=5e-4 for 80 epochs yielded a Silhouette of 0.4592. These experiments suggest diminishing returns on this small dataset for further Conv-only tuning; multimodal MLP remains the strongest performer for this dataset.

6 Figures

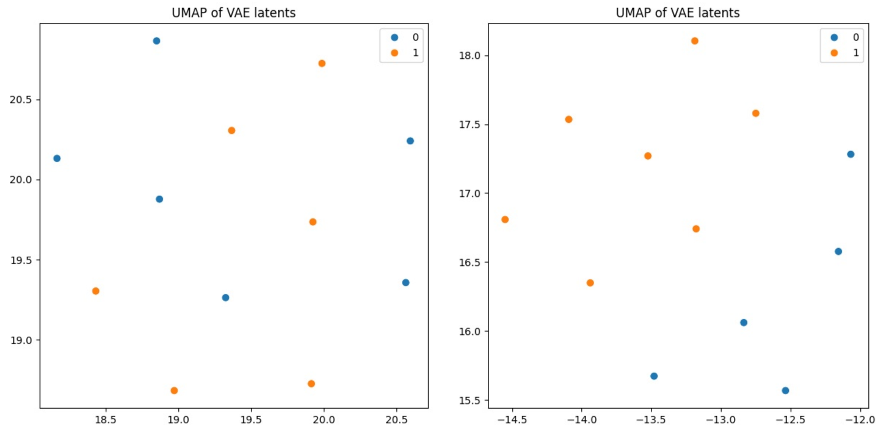

Figure 2: UMAP visualizations for (left) best ConvVAE run `conv_ld64_hc32` and (right) representative sweep run `conv_ld16_hc16`. Each shows cluster separation in the learned latent space.

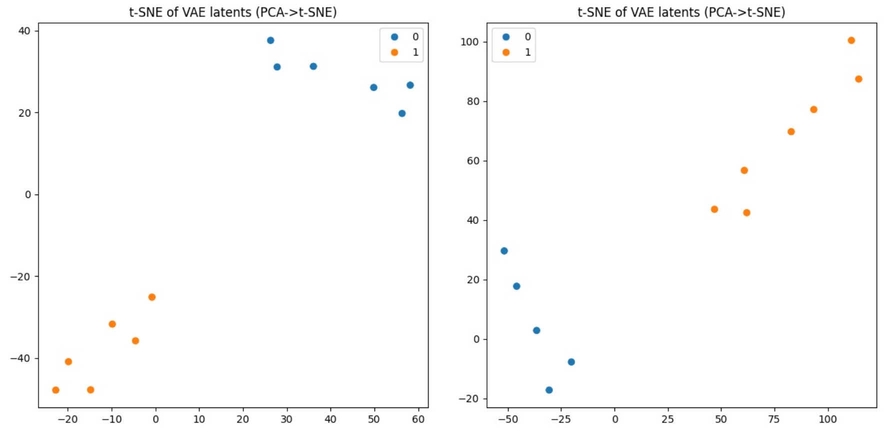

Figure 3: t-SNE visualizations for the same two runs, showing similar cluster structure under a different embedding.

7 Usage and Reproducibility

Run `python scripts/run_demo_pipeline.py` to reproduce the demo. To run on your dataset place audio and lyrics as described in `README.md` and run the pipeline steps.

8 Conclusion

This report demonstrates a working VAE-based pipeline for hybrid audio+lyrics clustering. On the provided Jamendo-style dataset the multimodal MLP VAE outperformed ConvVAE variants, though targeted ConvVAE sweeps produced moderate gains (best silhouette 0.5707). For submission-ready experiments, we recommend: (1) running ConvVAE on a larger and more diverse dataset or using data augmentation for spectrogram inputs, (2) exploring *beta*-VAE or conditional VAE variants, and (3) reporting stability across multiple folds and seeds. The repository includes scripts to reproduce all experiments and to generate the figures and metrics used in this report; an assembled HTML and PDF (via Playwright) are available under `results/`.