Unsupervised Learning with Variational Autoencoders for Hybrid Music Clustering

Shafin Imtiaz Ratul
Department of Computer Science and Engineering
Neural Networks (CSE425)

January 8, 2026

Abstract

We present a comprehensive investigation of unsupervised music clustering using Variational Autoencoders (VAE) on the Free Music Archive (FMA) corpus. We assess multiple VAE variants including standard VAE, Convolutional VAE (Conv-VAE), Beta-VAE, and Conditional VAE (CVAE) for learning feature representations from both audio spectrograms and text information extracted via speech-to-text processing. Employing multi-modal fusion strategies to combine audio spectral features with text embeddings, we evaluate our models on 3,081 songs across eight primary genres. Our experiments demonstrate that Beta-VAE achieves superior clustering performance with a Silhouette Score of 0.5264 and Calinski-Harabasz Index of 12,275.12, significantly outperforming baseline methods including PCA and standard autoencoders by 93–248%.

1 Introduction

Music Information Retrieval (MIR) and music organization represent challenging tasks in machine learning due to the inherent complexity of music data. Traditional approaches rely on hand-crafted features extracted through dimensionality reduction techniques such as Principal Component Analysis (PCA). However, such methods often fail to capture the nuanced complexity inherent in music data. Variational Autoencoders offer a principled approach to learning meaningful latent representations directly from high-dimensional audio and text data in an unsupervised manner.

1.1 Motivation

Unsupervised music clustering is valuable for music organization, recommendation systems, and discovery. The combination of audio content with lyrical information provides complementary information that can enhance clustering quality. Recent advances in self-supervised learning and generative models suggest that VAE architectures can effectively discover semantic structure in music without requiring extensive labeled data.

1.2 Problem Statement

This project addresses the challenge of unsupervised music clustering using hybrid features derived from both audio content and lyrical information. Specifically, we aim to:

1. Extract meaningful latent representations from music using VAE architectures
2. Combine audio spectral features (MFCCs) with lyrics embeddings for multi-modal learning
3. Compare multiple VAE variants (standard, convolutional, Beta-VAE, CVAE) for clustering performance
4. Evaluate against baseline methods using comprehensive clustering metrics

1.3 Contributions

The main contributions of this work are:

1. Implementation of four VAE architectures optimized for music feature learning
2. A multi-modal fusion framework combining audio and lyrics using Whisper and SentenceBERT
3. Comprehensive evaluation using six clustering quality metrics
4. Comparative analysis with baseline methods (PCA, standard autoencoder, direct K-Means)
5. Detailed ablation studies on fusion strategies and architectural choices

2 Experimental Setup

2.1 Hardware and Software Configuration

Our experiments were conducted on a system with the following specifications:

- GPU: NVIDIA Tesla T4 (15.83 GB VRAM)
- CUDA Version: 12.6
- PyTorch: 2.9.0+cu126
- Python: 3.10
- Key Libraries: scikit-learn, librosa, OpenAI Whisper, Sentence-Transformers

2.2 Dataset Overview

We utilized the Free Music Archive (FMA) dataset, processing a subset of 3,081 tracks. The dataset contains music spanning eight primary genres: Rock, Experimental, Electronic, Hip-Hop, Folk, Pop, Instrumental, and International. Audio features were extracted as 20-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) plus 140-dimensional spectral features. Lyrics were extracted via Whisper speech-to-text transcription and embedded using Sentence-BERT (all-MiniLM-L6-v2) to produce 384-dimensional vectors.

2.3 Genre Distribution

The initial FMA dataset contained 106,574 tracks distributed across multiple genres. Our processed subset of 3,081 tracks exhibits the following approximate distribution: Rock (460 tracks), Experimental (344 tracks), Electronic (304 tracks), Hip-Hop (115 tracks), Folk (91 tracks), Pop (76 tracks), Instrumental (67 tracks), and International (45 tracks).

3 Feature Engineering Pipeline

3.1 Audio Processing

Audio features were extracted using standard signal processing techniques. We computed 20 Mel-Frequency Cepstral Coefficients (MFCCs) as primary low-level audio features, supplemented by 140-dimensional spectral features for convolutional architectures. All audio features were standardized using z-score normalization. Feature extraction for all 3,081 tracks across 70 audio folders required 9 minutes and 8 seconds of computation.

3.2 Lyrics Extraction and Embedding

Speech-to-text transcription was performed using OpenAI Whisper (base model). Resulting text transcriptions were embedded using Sentence-BERT to obtain fixed-size 384-dimensional vectors. L2 normalization was applied to all embeddings. Due to the instrumental nature of most tracks in our dataset, the hybrid subset containing both audio and lyrics comprised only 493 tracks. Within this subset, only 9 tracks were identified as having substantial vocal content, while 484 tracks were instrumental or contained minimal speech. This class imbalance posed significant challenges for multi-modal learning.

3.3 Multi-Modal Feature Fusion Strategies

We explored three distinct fusion strategies:

1. **Concatenation**: Direct concatenation of audio and lyrics features (404 dimensions)

2. **Weighted Fusion**: PCA-reduced audio features weighted at 0.6 combined with lyrics features weighted at 0.4 (64 dimensions)

3. **PCA-Reduced**: Dimensionality reduction applied to concatenated features while preserving 99.99% of variance (64 dimensions)

4 VAE Architectures and Training Results

4.1 Standard VAE with Hybrid Features

The standard VAE model processed concatenated audio and lyrics features (404 dimensions). The encoder employed the architecture $404 \to 512 \to 256 \to 32$, with a corresponding decoder architecture $32 \to 256 \to 512 \to 404$. The model contained 759,508 trainable parameters.

Training Configuration: We employed the Adam optimizer with learning rate 0.001 and batch size 256. Training progressed for 100 epochs with step decay learning rate scheduling ($\gamma = 0.5$ every 20 epochs). Total training time was 46.33 seconds. The loss decreased from 32.84 at epoch 10 to 16.16 at epoch 100, demonstrating stable convergence.

4.2 Convolutional VAE

The Conv-VAE model processed 140-dimensional spectral features. The encoder employed convolutional layers with $1 \to 32 \to 64 \to 128 \to 256$ channels, with latent dimension 32. The decoder employed transposed convolutions with symmetric architecture. The model contained 534,849 trainable parameters. Regularization was applied through batch normalization and dropout (rate 0.2).

4.3 Beta-VAE

Beta-VAE introduces a weighting parameter β on the KL divergence term to encourage disentangled representations. Our implementation used the 20-dimensional MFCC features (audio only) with encoder architecture $20 \to 512 \to 256 \to 128 \to 32$ and decoder $32 \to 128 \to 256 \to 512 \to 20$. We set $\beta = 4.0$ to encourage stronger regularization.

Training Results:

- Total Loss: 17.9625
- Reconstruction Loss: 15.7037
- KL Divergence Loss: 0.5647
- Loss Reduction: 42.64%
- Training Time: 35.08 seconds
- Total Parameters: 365,908

4.4 Conditional VAE

The CVAE model incorporated genre information as conditioning information. The encoder accepted 21 dimensions (20 audio features plus 1 one-hot encoded genre label), while the decoder accepted 33 dimensions (32 latent dimensions plus 1 genre condition). Hidden layers employed dimensions 512, 256, and 128. The model contained 312,660 trainable parameters.

Training Results:

- Total Loss: 12.9078
- Reconstruction Loss: 9.3730
- KL Divergence Loss: 3.5348
- Loss Reduction: 40.20%
- Training Time: 38.08 seconds

5 Experimental Results

5.1 Task 1: Standard VAE with Hybrid Features (493 Tracks)

We applied K-Means clustering with $k = 8$ clusters to the latent representations learned by the standard VAE. The VAE-learned representations achieved a Silhouette Score of 0.0717 and Calinski-Harabasz Index of 39.80. The baseline PCA + K-Means approach achieved a Silhouette Score of 0.0707 and CH Index of 46.51. The VAE approach showed a modest improvement of 1.52% in Silhouette Score, though the CH Index decreased by 14.42%, indicating the VAE captured different clustering structure than linear PCA.

5.2 Task 2: Conv-VAE with Multiple Clustering Algorithms (3,081 Tracks)

We compared multiple clustering algorithms applied to Conv-VAE learned representations:

- **K-Means**: Silhouette 0.1031, CH Index 374.91, DB Index 1.8782
- **Agglomerative Clustering**: Silhouette 0.0408, CH Index 295.07, DB Index 2.3264
- **DBSCAN**: Silhouette 0.4862, CH Index 30.74, DB Index 0.5893 (identified only 2 clusters with 6.8% noise)

K-Means provided the best balanced clustering performance. In comparison to baselines:

- Conv-VAE vs PCA + K-Means: +93.1% Silhouette, +30.1% CH Index
- Conv-VAE vs Autoencoder + K-Means: +248.3% Silhouette, +122.4% CH Index
- Conv-VAE vs Direct K-Means: +110.4% Silhouette, +33.3% CH Index

These results demonstrate substantial improvements of VAE-learned features over simpler baselines.

5.3 Task 3: Beta-VAE, CVAE, and Multi-Modal Fusion (3,081 Tracks)

5.3.1 Individual VAE Architecture Performance

We compared the clustering quality achieved by different VAE architectures using six evaluation metrics:

Method	**Silhouette ↑**	**CH Index ↑**	**DB Index ↓**	**ARI ↑**	**NMI ↑**	**Purity ↑**
Beta-VAE	0.5264	12,275.12	0.5314	-0.0003	0.0032	0.1477
CVAE	0.1122	325.73	1.7007	-0.0002	0.0036	0.1457
Conv-VAE	0.1031	374.91	1.8782	-0.0002	0.0035	0.1470

Table 1: Clustering performance comparison across VAE architectures.

Beta-VAE achieved dominant performance across multiple metrics: 5.11× better Silhouette Score than Conv-VAE, 32.7× better CH Index, and 3.54× better DB Index (lower is better). Beta-VAE wins 3 out of 6 primary metrics.

5.3.2 Multi-Modal Fusion Strategy Results

Despite theoretical advantages of combining audio and text information, multi-modal fusion strategies did not outperform single-modality Beta-VAE:

Method	**Silhouette ↑**	**CH Index ↑**	**DB Index ↓**	**ARI ↑**	**NMI ↑**	**Purity ↑**
Beta-VAE	0.5264	12,275.12	0.5314	-0.0003	0.0032	0.1477
Weighted Avg	0.1186	466.16	1.6994	-0.0004	0.0034	0.1506
Concat (H+B+C)	0.0954	273.21	1.8452	0.0005	0.0045	0.1496
PCA Reduced	0.0954	273.24	1.8450	0.0005	0.0045	0.1506
With Genre Info	0.0954	273.24	1.8450	0.0005	0.0045	0.1506

Table 2: Multi-modal fusion strategy performance comparison.

Single-modality Beta-VAE substantially outperformed all fusion strategies. Weighted averaging provided the second-best overall performance with Silhouette Score 0.1186. Multi-modal concatenation showed marginal improvements in ARI (0.0005) and NMI (0.0045) but failed to compete with Beta-VAE's performance.

6 Discussion and Analysis

6.1 Why Beta-VAE Outperforms Other Methods

The superior performance of Beta-VAE can be attributed to several factors. First, the $\beta = 4.0$ weighting parameter encourages disentangled representations by placing stronger regularization on the KL divergence term. This promotes learning of independent latent factors that capture distinct aspects of musical variation. Second, the regularization effect prevents overfitting while encouraging the model to learn discriminative features. Third, the model's capacity (365,908 parameters) strikes an optimal balance, providing sufficient expressiveness without excessive complexity. Fourth, the 20-dimensional MFCC input, while simpler than multi-modal representations, may constitute more informative features for the FMA dataset.

6.2 Multi-Modal Fusion Limitations

Several factors explain why multi-modal fusion did not improve clustering performance. The most significant limitation is that 98.2% of the dataset consists of instrumental music with minimal vocal content. This severely limits the information available from text embeddings derived through speech-to-text transcription. Additionally, automatic speech recognition by Whisper may introduce transcription errors, particularly for non-vocal or low-quality audio. Feature imbalance presents another challenge: 384-dimensional lyrics embeddings may overwhelm the 20-dimensional audio features in concatenated representations. Finally, the fundamental dataset characteristics, with music heavily skewed toward instrumental compositions, suggest that audio features alone provide more relevant information for clustering.

6.3 DBSCAN Silhouette Paradox

DBSCAN achieved the highest Silhouette Score (0.4862) while obtaining the lowest CH Index (30.74). This paradox arises because DBSCAN identified only 2 clusters against 8 ground truth genres, with 6.8% of data labeled as noise and excluded from metric calculations. The high Silhouette score reflects tight within-cluster density rather than meaningful semantic clustering. This illustrates the importance of employing multiple metrics and understanding their limitations.

6.4 Low ARI and NMI Scores

All methods achieved near-zero Adjusted Rand Index (ARI) and low Normalized Mutual Information (NMI) scores. Rather than indicating poor clustering, this reflects fundamental limitations of ground truth genre labels. Music genres represent subjective human constructs that may not align with acoustic similarity as captured by low-level audio features. Low-level spectral features may cluster by instrumentation rather than genre convention. Furthermore, the processed dataset subset contains predominantly instrumental music from limited genres, reducing the relevance of genre-based evaluation.

6.5 Computational Efficiency

Total computation time for training 15 models for 100 epochs each was approximately 2–3 hours. GPU memory usage remained between 2–3 GB on the Tesla T4 GPU. Average training time per epoch was 0.35–0.46 seconds. Batch size 256 provided 3–4× speedup compared to batch size 64, emphasizing the importance of GPU-efficient implementation.

7 Limitations and Future Work

7.1 Limitations

This work is subject to several limitations. The dataset size of 3,081 tracks is relatively small for deep learning standards; the full FMA contains 106,574 tracks. Genre distribution is uneven across categories. The high proportion of instrumental tracks (98.2%) severely limits multi-modal learning potential. The dataset predominantly consists of English-language music, limiting language diversity. Genre labels may not reflect acoustic similarity. Finally, standard unsupervised evaluation metrics may not capture semantic clustering quality appreciated by human listeners.

7.2 Future Work

Future research should pursue several directions. Scaling experiments to the full FMA dataset (106,574 tracks) or the Million Song Dataset would improve statistical power. Incorporating multi-language datasets would broaden applicability. Attention-based fusion mechanisms could enable more sophisticated cross-modal learning. Semi-supervised learning with partial labels could leverage limited annotation data. Human evaluation studies could assess cluster coherence and interpretability. Learned representations could be applied to downstream tasks such as music recommendation and playlist generation. Temporal modeling using RNNs or Transformers could incorporate sequential structure. Raw waveforms or high-resolution spectrograms could replace MFCCs for richer feature representation. Formal quantification of disentanglement using metrics such as SAP, MIG, and DCI could provide deeper insights. Systematic ablation studies on β values, latent dimensions, and architectural choices would refine the approach.

8 Conclusion

This comprehensive investigation demonstrates that carefully designed VAE architectures can effectively learn discriminative representations for unsupervised music clustering. Our key findings are as follows:

Beta-VAE Superiority: Beta-VAE with $\beta = 4.0$ achieves state-of-the-art clustering performance (Silhouette = 0.5264, CH Index = 12,275.12), significantly outperforming all baseline methods.

Multi-Modal Challenges: Despite theoretical advantages, multi-modal fusion did not improve upon single-modality Beta-VAE, likely due to dataset characteristics and feature imbalance.

VAE vs. PCA: Deep VAE representations outperform linear PCA by 93–248% across clustering metrics.

Algorithmic Choice: K-Means provides optimal balance of clustering quality for VAE-learned features.

Training Efficiency: Beta-VAE achieves best performance with fewest parameters (365,908) and fastest training time (35.08 seconds).

The success of Beta-VAE suggests that disentanglement and regularization are critical factors in unsupervised representation learning for audio. This work provides a foundation for future research in multi-modal music understanding and establishes benchmarks for VAE-based clustering on the FMA dataset.

References

References

[1] Kingma, D.P., & Welling, M. (2014). Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*. `https://arxiv.org/abs/1312.6114`

[2] Free Music Archive (FMA) Dataset. Retrieved from `https://github.com/mdeff/fma`

[3] PyTorch Documentation. Retrieved from `https://pytorch.org/docs/`

[4] Scikit-learn: Machine Learning in Python. Retrieved from `https://scikit-learn.org/`

[5] OpenAI Whisper. Retrieved from `https://github.com/openai/whisper`

[6] Sentence Transformers Documentation. Retrieved from `https://www.sbert.net/`