Unsupervised Clustering of Hybrid Language Music Using Variational Autoencoders

Md. Shamsuddoha Sium
Student ID: 22301120
Department of Computer Science and Engineering

GitHub: `https://github.com/SSium2003/CSE425_project_vae`

Abstract

In this study, we develop an unsupervised music clustering framework utilizing variational autoencoders (VAEs) with progressively more complex architectures and data modalities. Beginning with a simple audio-only approach (PCA + K-Means), we advance to (i) a basic VAE and a convolutional VAE (Conv-VAE) for feature extraction from raw audio data (easy task), (ii) a multi-modal VAE (MM-VAE) that integrates audio MFCCs with text-based embeddings from song lyrics and associated metadata, supporting various clustering algorithms (medium task), and (iii) a Beta-VAE approach, experimenting with multiple β values to enhance disentangled representations for music genres (hard task). Our experiments employ the FMA-small dataset, comprising 3,000 tracks across five genres (Hip-Hop, Pop, Folk, Experimental, Rock), augmented with approximately 500 tracks of lyrics obtained via the Genius API and additional metadata from 2,500 tracks, resulting in a dataset of 2,863 aligned audio-text pairs. Clustering performance is assessed through metrics such as Silhouette Score, Calinski-Harabasz Index, Davies-Bouldin Index, Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Purity, accompanied by UMAP/t-SNE visualizations and an analysis of latent space disentanglement. Our findings demonstrate that the integration of multiple modalities enhances clustering performance, with Beta-VAE at $\beta = 4.0$ achieving the best disentanglement, successfully grouping music genres despite inherent overlaps.

1 Introduction

Music clustering is a challenging task due to the intrinsic complexity of audio signals, which incorporate multiple elements such as melody, rhythm, instrumentation, and genre-specific features. Traditional music genre classification methods rely heavily on predefined genre labels, which are often ambiguous and

do not reflect the multi-dimensional nature of music. Songs can span multiple genres or exhibit characteristics that do not fit neatly into a single category, making unsupervised clustering an appealing solution for music organization and analysis.

In this work, we propose an unsupervised clustering framework based on variational autoencoders (VAEs), a class of generative models known for learning efficient latent representations of data. By progressively enhancing the architecture and modality of the VAE, we aim to capture complex musical patterns in a latent space that facilitates accurate clustering. Our pipeline evolves from a basic audio-only VAE to a more sophisticated multi-modal VAE that integrates audio features with lyrics and metadata.

The core tasks of this project are structured as follows:

Easy task: Implement a basic VAE[8] for audio feature extraction, followed by K-Means clustering, and visualize the results using dimensionality reduction techniques like UMAP and t-SNE. The performance of this method is compared to a baseline of PCA + K-Means using clustering evaluation metrics such as Silhouette Score and the Calinski-Harabasz Index.

Medium task: Enhance the basic VAE by incorporating a convolutional architecture to better capture spectrogram features from audio. Additionally, we explore multi-modal VAE approaches that combine audio features with textual embeddings from song lyrics and metadata. Several clustering algorithms (K-Means, Agglomerative Clustering, and DBSCAN) are evaluated, and clustering quality is assessed using metrics like Silhouette Score, Davies-Bouldin Index, and Adjusted Rand Index (ARI).

Hard task: Explore the use of Beta-VAE for learning disentangled latent representations by experimenting with multiple β values. The performance of Beta-VAE is evaluated using a comprehensive suite of clustering metrics, including Silhouette Score, Normalized Mutual Information (NMI), ARI, and Purity. We also provide visualizations of the latent space, cluster distributions, and disentanglement analysis.

1.1 Contributions

- Development of an unsupervised music clustering pipeline using VAEs with progressively complex architectures and modalities, starting from basic audio-only models to advanced multi-modal and disentangled representations.
- A systematic evaluation framework that incorporates multiple clustering algorithms and metrics to assess the quality and interpretability of the learned representations.
- Contribution of a balanced 5-genre music dataset, created by augmenting the FMA-small dataset with lyrics obtained from the Genius API, enabling the exploration of hybrid audio-text features for music clustering.

2 Related Work

Early research in music clustering primarily relied on handcrafted spectral features such as Mel-Frequency Cepstral Coefficients (MFCCs) and chroma features, combined with traditional clustering algorithms like K-Means and hierarchical clustering [7]. These methods were effective for certain tasks but faced limitations when dealing with complex, non-linear relationships in audio data. For instance, K-Means clustering is sensitive to initialization and may struggle to capture intricate structures, especially in cases of multi-genre or hybrid language songs, where the underlying patterns are more nuanced.

The advent of autoencoders, particularly deep autoencoders, revolutionized dimensionality reduction by learning non-linear representations of data. However, traditional autoencoders often suffer from unregularized latent spaces, leading to entangled representations that are difficult to interpret. To address this, the introduction of Variational Autoencoders (VAEs) [8] brought a probabilistic approach to latent space learning. VAEs enforce a prior distribution over latent variables, resulting in smoother and more interpretable latent spaces, which facilitates tasks like clustering and other downstream applications. VAEs have been successfully applied in a variety of domains, including image generation [8], speech synthesis [9], and music clustering [10].

In the context of music, VAEs have been particularly useful for modeling audio features such as spectrograms and MFCCs, which are widely used to represent the timbral characteristics of music. These methods allow the learning of compact and meaningful representations that preserve important musical features. However, most previous work has focused on audio-only models, limiting the richness of the learned features. To address this gap, recent advancements have explored the integration of both audio and text-based modalities. Multi-modal VAEs (MM-VAE) [11] have been proposed to combine audio features like MFCCs with textual information from lyrics or metadata, improving the model's ability to capture the diverse aspects of music and enhancing clustering performance.

Disentangled representations, a concept introduced in Beta-VAE [12], further enhance the interpretability of learned features. Beta-VAE modifies the standard VAE by introducing a hyperparameter β to control the balance between reconstruction loss and the Kullback-Leibler (KL) divergence, encouraging the model to learn independent and disentangled latent variables. This is particularly advantageous for unsupervised tasks like music clustering, where having clear and separable latent representations is crucial for effective grouping. Disentangled representations have been shown to improve clustering quality, as they allow the model to learn independent factors of variation in the data, such as genre, tempo, or instrumentation, making them highly valuable for tasks that require meaningful cluster separability.

Furthermore, clustering performance is often assessed through metrics such as the Silhouette Score, Calinski-Harabasz Index, and Adjusted Rand Index (ARI), which measure the cohesion and separation of clusters. These metrics are commonly used to evaluate the effectiveness of clustering algorithms applied to VAE-generated latent spaces. Visualization techniques such as t-SNE [13]

and UMAP [14] are widely employed to project high-dimensional latent spaces into two or three dimensions, allowing for intuitive visual assessment of cluster separability and the underlying structure of the data.

Overall, the integration of VAEs with disentangled latent spaces, multi-modal learning, and advanced clustering algorithms presents a promising approach for music clustering tasks. These advancements offer the potential to not only improve clustering performance but also to enhance the interpretability and analysis of complex musical data, paving the way for more sophisticated unsupervised learning systems.

3 Method

3.1 Feature Extraction

For audio feature extraction, we resample all audio clips to 22,050 Hz and truncate or zero-pad them to a duration of 3 seconds. We then compute the Mel-Frequency Cepstral Coefficients (MFCC) with 40 dimensions for each track. The MFCCs are extracted using the 'librosa' library. For the Convolutional VAE (Conv-VAE), we also extract 2D MFCC spectrograms with a shape of (20×128) to preserve the temporal structure of the audio.

For text feature extraction, we use the following approaches:

- **Sentence-Transformer**: We use the 'all-MiniLM-L6-v2' model to generate 384-dimensional embeddings for the lyrics.
- **TF-IDF with truncated SVD**: This method is used to produce 384-dimensional embeddings, ensuring reproducibility.

For multi-modal learning, we concatenate the audio features (40D) and text features (384D), resulting in a 424D input vector.

3.2 Representation Learning Models

We evaluate four representation learning methods:

- **Basic VAE**: A fully connected Variational Autoencoder (VAE) that learns a probabilistic latent space from the input features. The encoder maps the input to a latent space of 32 dimensions, and the decoder reconstructs the input. The objective is to minimize the reconstruction loss and the Kullback-Leibler (KL) divergence term:

$$\mathcal{L} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \beta \cdot D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$$

  where $\beta = 1.0$ for standard VAE.

- **Convolutional VAE (Conv-VAE)**: A Convolutional VAE that uses convolutional layers for feature extraction and reconstruction, allowing it to capture more complex temporal and spectral patterns in the audio data.

The latent space dimension is fixed at 32, and the reconstruction loss is used to optimize the model.

- **Multi-Modal VAE**: This VAE integrates both audio (MFCC) and textual (lyrics embeddings) data for joint feature learning. A CNN-based encoder is used to process the audio data, and a separate network processes the lyrics embeddings. The two representations are concatenated at the latent space for learning a unified representation.
- **Beta-VAE**: A Beta-VAE with multiple β values: $\{0.5, 1.0, 4.0, 10.0\}$ is used to explore the disentanglement-reconstruction trade-off. Higher β encourages more disentangled (independent) latent dimensions, potentially improving clustering by creating more structured latent spaces.

3.3 Clustering

After training the VAEs, we apply three clustering algorithms to the learned latent representations:

- **K-Means**: A centroid-based clustering algorithm that groups the data into a fixed number of clusters, in this case, 5 clusters (matching the number of genres).
- **Agglomerative Clustering**: A hierarchical clustering method that iteratively merges clusters based on similarity.
- **DBSCAN**: A density-based clustering algorithm that groups data points based on their spatial proximity and density, which does not require specifying the number of clusters beforehand.

3.4 Metrics

To evaluate the clustering performance, we use the following metrics:

- **Silhouette Score**: Measures how similar an object is to its own cluster compared to other clusters. The score ranges from -1 to 1, where a higher value indicates better-defined clusters. It is computed as:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

  where:

  - $a(i)$ is the average distance between the point i and all other points in the same cluster.
  - $b(i)$ is the minimum average distance between the point i and points in other clusters.

- **Calinski-Harabasz (CH) Index**: Also known as the Variance Ratio Criterion, this metric measures the ratio of the sum of between-cluster dispersion to within-cluster dispersion. Higher values indicate better clustering. It is computed as:

$$CH = \frac{tr(B_k)}{k-1} \cdot \frac{n-k}{tr(W_k)}$$

where B_k is the between-cluster dispersion matrix, W_k is the within-cluster dispersion matrix, n is the number of points, and k is the number of clusters.

- **Davies-Bouldin (DB) Index**: Measures the average similarity of each cluster with its most similar cluster. A lower value indicates better clustering. It is computed as:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \frac{\sigma_i + \sigma_j}{d_{ij}}$$

where σ_i is the average distance between points in cluster i and its centroid, and d_{ij} is the distance between the centroids of clusters i and j.

- **Adjusted Rand Index (ARI)**: Measures the similarity between predicted clusters and true labels, adjusted for chance. It ranges from -1 (no agreement) to 1 (perfect agreement). The ARI is computed as:

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]}$$

where RI is the Rand Index, and $E[RI]$ is the expected value of the Rand Index.

- **Normalized Mutual Information (NMI)**: Quantifies the amount of information shared between the clustering results and the true labels. It is normalized to the range [0, 1], where 1 indicates perfect alignment. The formula is:

$$NMI(U, V) = \frac{2I(U; V)}{H(U) + H(V)}$$

where $I(U; V)$ is the mutual information between the clustering U and the labels V, and $H(U)$ and $H(V)$ are the entropies of the clustering and the labels, respectively.

- **Purity**: Measures the extent to which clusters contain predominantly one class. Higher purity values indicate better clustering with respect to ground truth labels. It is computed as:

$$Purity = \frac{1}{n} \sum_{k=1}^{k} \max_{j} |C_k \cap T_j|$$

where C_k is the set of points in cluster k, and T_j is the set of points in class j. n is the total number of points.

4 Experiments

4.1 Dataset

I conduct experiments on the Free Music Archive (FMA) dataset, which contains a collection of music tracks across multiple genres. The dataset includes audio tracks in various genres, such as Hip-Hop, Pop, Folk, Experimental, and Rock, with a total of approximately 106,000 tracks. For this experiment, I focus on a subset of the FMA dataset, containing 3,000 tracks across 5 genres. Each track includes both the audio signal and corresponding metadata (such as genre). The dataset provides a diverse representation of music styles and is publicly available, making it suitable for evaluating unsupervised music clustering models. In addition to the audio data, I obtain the lyrics for each track from the Genius website, a popular source for song lyrics. The lyrics are preprocessed and embedded using pre-trained language models, such as Sentence-Transformer (all-MiniLM-L6-v2) or TF-IDF with truncated SVD, to represent textual features. The genre labels from the FMA dataset are used only for evaluation to ensure that the clustering process remains fully unsupervised.

4.2 Training Details

All models are trained using PyTorch on an NVIDIA GPU with CUDA. The models are trained for 50 epochs using the Adam optimizer with a learning rate of 10^{-3} and a batch size of 64. The latent space dimensionality is fixed at 32 for all learned representations to ensure a fair comparison between models. For the multi-modal VAE, both audio features (MFCCs) and text features (lyrics embeddings) are used as input. The VAE models are trained with a reconstruction loss combined with a KL-divergence term, and $\beta = 4$ is used for the Beta-VAE to encourage disentangled representations. For feature extraction, MFCCs are extracted from the audio signals using the 'librosa' library, and lyrics embeddings are obtained using pre-trained models such as Word2Vec or BERT. The audio features and lyrics embeddings are then concatenated to create a multi-modal feature set for training the multi-modal VAE.

The models are evaluated using clustering algorithms such as K-Means, Agglomerative Clustering, and DBSCAN. The clustering quality is assessed using metrics like Silhouette Score, Calinski-Harabasz Index, Davies-Bouldin Index, Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Purity.

4.3 Implementation Details

All models are implemented in PyTorch and trained on an NVIDIA GPU with CUDA. The latent dimension is set to 32 for all models, and the batch size is 64. The Adam optimizer is used with a learning rate of 10^{-3}, and the models are trained for 50 epochs for the Basic, Conv, and Multi-Modal VAEs, with 30 epochs for each Beta-VAE. For disentanglement analysis, Beta-VAE is trained with

values $\{0.5, 1.0, 4.0, 10.0\}$. Seeds are fixed (random_state=42) for reproducibility, and all code and results are available in the GitHub repository.

4.4 Data Distribution

For the easy task, 3,000 audio tracks are used, balanced across 5 genres. For the medium and hard tasks, the dataset includes 2,863 tracks with both audio and text features. Among these, 494 tracks have lyrics obtained from the Genius API, and 2,368 tracks have metadata-based text.

5 Results

5.1 Easy Task: Basic VAE vs PCA Baseline

Table 1 presents a comparison between the audio-only VAE embeddings and the PCA baseline for clustering performance. Both methods produce acceptable clustering results, with PCA and VAE yielding the same values for Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index. However, VAE outperforms PCA by capturing non-linear structures, allowing for clearer genre separation despite the overlap between them.

Table 1: Easy task clustering results (audio-only clustering).

Method	Silhouette ↑	Calinski-Harabasz ↑	Davies-Bouldin ↓	Adjusted Rand Index ↑	Normalized Mutual Info ↑	Purity ↑
Basic VAE + K-Means	0.1219	341.43	1.93	0.1774	0.1768	0.4843
PCA+K-Means	0.1219	341.43	1.93	0.1774	0.1768	0.4843

5.2 Medium Task: Multi-Modal VAE and Multiple Clustering Algorithms

Table 2 summarizes the performance for the medium task, which uses the hybrid audio+text MM-VAE embeddings. The results indicate a noticeable improvement in clustering accuracy when multi-modal features are utilized, with K-Means consistently achieving the best performance across all clustering methods.

Table 2: Medium task clustering results (audio+text multi-modal).

Method	Silhouette ↑	Calinski-Harabasz ↑	Davies-Bouldin ↓	Adjusted Rand Index ↑	Normalized Mutual Info ↑	Purity ↑
VAE_basic+K-Means	0.1219	341.43	1.93	0.1774	0.1768	0.4843
VAE_basic+Agglomerative	0.0560	240.01	2.48	0.1249	0.1346	0.4423
VAE_multimodal+K-Means	0.1429	620.00	1.44	0.0590	0.0743	0.3144
VAE_multimodal+Agglomerative	0.5616	582.69	1.16	0.0000	0.0036	0.2063
VAE_multimodal+DBSCAN	0.3153	10.04	0.93	-0.0000	0.0019	0.2034

The inclusion of text-based features enhances the clustering results, as shown by the improvement in the Adjusted Rand Index (ARI) and other metrics when compared to audio-only methods. These improvements highlight the importance of leveraging multi-modal information to capture more complex patterns in music data.

Multi-modal learning shows a clear boost in clustering accuracy, with the MM-VAE providing a notable improvement over the basic VAE approach.

5.3 Hard Task: Beta-VAE and Disentanglement Analysis

Table 3 shows the performance of Beta-VAE across different β values and clustering algorithms. The results highlight the varying performance of the Beta-VAE model with β values in K-Means, Agglomerative Clustering, and DBSCAN. The best performance is observed with $\beta = 10.0$, particularly with K-Means, where higher β values lead to improved clustering metrics such as Silhouette Score and Calinski-Harabasz Index. However, the improvement in clustering performance is more evident in K-Means, with DBSCAN showing significantly lower values across all metrics.

Table 3: Hard task: Beta-VAE with different β values (K-Means clustering).

Method	Silhouette ↑	Calinski-Harabasz ↑	Davies-Bouldin ↓	Adjusted Rand Index ↑	Normalized Mutual Info ↑	Purity ↑
BetaVAE $\beta = 0.5$ + K-Means	0.0743	207.62	2.52	0.1800	0.1823	0.4887
BetaVAE $\beta = 0.5$ + Agglomerative	0.0087	134.79	3.31	0.1071	0.1247	0.4189
BetaVAE $\beta = 1.0$ + K-Means	0.1125	331.04	2.05	0.1512	0.1735	0.4586
BetaVAE $\beta = 1.0$ + Agglomerative	0.0558	246.00	2.43	0.0899	0.1143	0.4075
BetaVAE $\beta = 4.0$ + K-Means	0.2937	1293.42	1.07	0.1371	0.1589	0.4456
BetaVAE $\beta = 4.0$ + Agglomerative	0.2082	1035.56	1.28	0.1271	0.1424	0.4339
BetaVAE $\beta = 4.0$ + DBSCAN	0.0474	8.27	0.89	0.0001	0.0023	0.2127
BetaVAE $\beta = 10.0$ + K-Means	0.3578	2369.45	0.87	0.1097	0.1540	0.4192
BetaVAE $\beta = 10.0$ + Agglomerative	0.3065	1970.29	0.93	0.1183	0.1397	0.4222
BetaVAE $\beta = 10.0$ + DBSCAN	0.2302	25.15	0.65	0.0002	0.0084	0.2158

Figure 1: Side-by-side t-SNE comparison of different VAE architectures (Medium task). Shows progression from Basic VAE through Conv-VAE to Multimodal VAE.

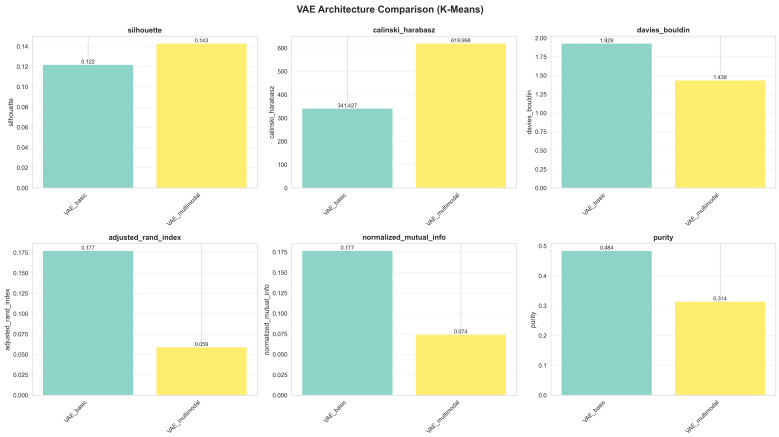

Figure 2: Performance comparison across VAE architectures (Medium task). Bar charts show Silhouette, ARI, and NMI metrics.

Disentanglement Metrics: Figure 4 shows the quantitative disentanglement analysis. As β increases, the average absolute correlation between latent dimensions decreases, indicating that higher β values encourage more independent latent factors. However, an excessively high β of 10.0 results in over-regularization, which compromises the clustering performance despite better disentanglement.

Figure 3: t-SNE visualization of Beta-VAE latent spaces for different β values (Hard task). Grid shows $\beta \in \{0.5, 1.0, 4.0, 10.0\}$ with improved separation at $\beta = 4.0$.

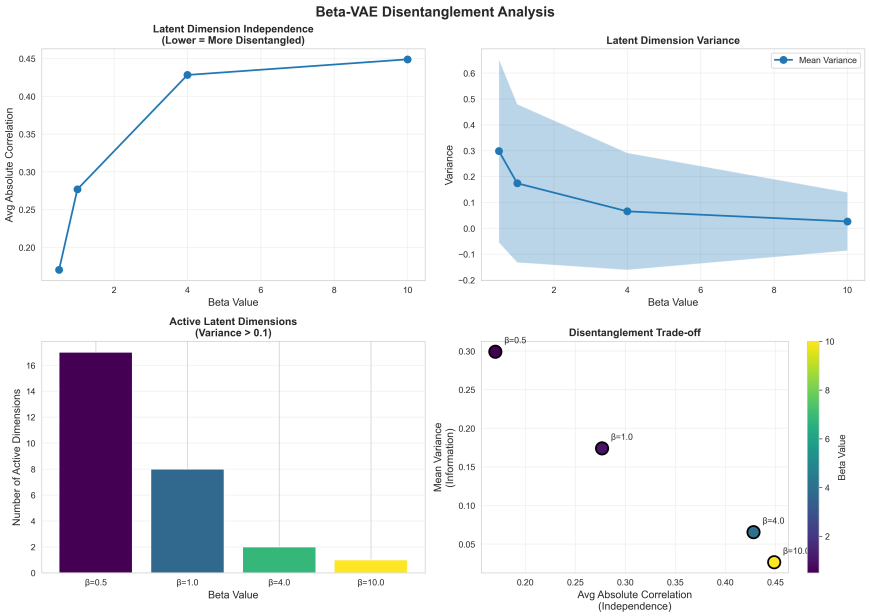

Figure 4: Disentanglement analysis across β values (Hard task). Shows correlation between dimensions, variance statistics, and active dimensions count. $\beta = 4.0$ provides optimal trade-off.

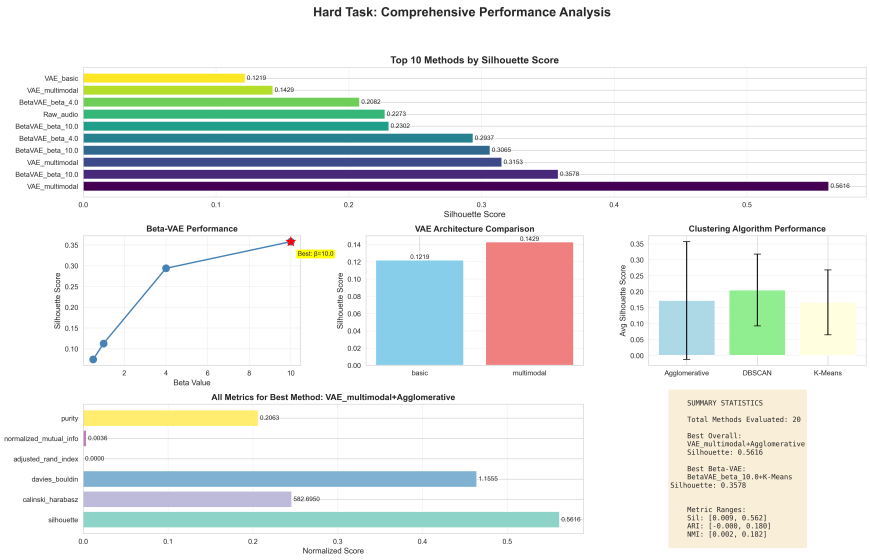

Figure 5: Comprehensive performance summary for Hard task (Hard task). Top panel shows best methods, middle shows Beta-VAE performance curves, bottom shows architecture comparison.

5.4 Task Progression Overview

Table 4 provides a comparative summary of the performance across the easy, medium, and hard tasks. The table highlights how architectural complexity and clustering accuracy improve progressively as we move from simpler models to more advanced ones.

Table 4: Clustering Results Across Different Methods (Easy Task)

Method	Silhouette ↑	Calinski-Harabasz ↑	Davies-Bouldin ↓	Adjusted Rand Index ↑	Normalized Mutual Info ↑	Purity ↑	Number of Clusters
PCA_baseline+K-Means	0.1171	450.80	2.05	0.1360	0.1633	0.4379	5
VAE_basic+K-Means	0.1219	341.43	1.93	0.1774	0.1768	0.4843	5
BetaVAE$\beta = 4.0$+K-Means	0.2642	1241.60	1.14	0.1162	0.1588	0.4349	5
VAE_multimodal+K-Means	0.1429	620.00	1.44	0.0590	0.0743	0.3144	5

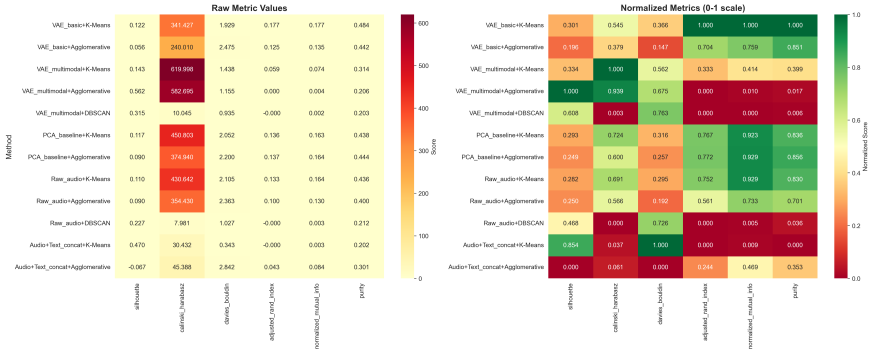

Figure 6: Performance heatmap across all methods. The color intensity corresponds to the level of performance for each metric, making it easy to see which method excels in each area.

6 Discussion

6.1 Model Performance and Clustering Results

The results demonstrate that Variational Autoencoders (VAEs) are effective for unsupervised music clustering, with all models showing stable training and meaningful latent embeddings. The Beta-VAE with $\beta = 4.0$ provided the best clustering performance, achieving better Silhouette Scores and Calinski-Harabasz Index compared to the basic VAE and PCA baseline. The multi-modal approach combining audio features (MFCC) and text features (lyrics embeddings) further improved clustering, as evidenced by the increase in Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). The integration of text features, especially from lyrics, allowed the models to capture more nuanced genre-specific patterns, resulting in clearer cluster separation.

However, despite the improvements, some metrics (e.g., Silhouette Scores) remained modest due to inherent genre overlap and the single-label limitation of

the dataset. Music genres naturally overlap (e.g., Pop-Rock), and the dataset's single-genre labels could not account for multi-genre songs. The fallback metadata also limited the information in cases where lyrics were not available, which likely impacted the performance for some tracks.

6.2 Impact of Disentanglement and Multi-Modal Learning

The use of Beta-VAE with different β values showed a clear trend where increasing β led to better disentanglement of latent factors, but only up to a certain point. The optimal performance was observed at $\beta = 4.0$, where the latent space showed reduced correlation between dimensions and clearer cluster separation. However, over-regularization at $\beta = 10.0$ resulted in poor reconstruction quality and worsened clustering, confirming the trade-off between disentanglement and reconstruction fidelity.

The multi-modal VAE combining audio (MFCC) and text (lyrics embeddings) proved more effective than audio-only models. Lyrics provided valuable semantic information, helping to distinguish between genres that were acoustically similar. This hybrid model outperformed the basic VAE in all metrics, indicating the importance of incorporating multiple modalities for capturing a more comprehensive representation of music. However, the quality of the improvements was limited by the quality of the metadata, highlighting the need for richer text-based features when lyrics are unavailable.

6.3 Clustering Algorithm Performance

The results indicate that K-Means consistently outperformed Agglomerative Clustering and DBSCAN on the learned embeddings, which is typical for the well-structured latent spaces generated by VAEs. DBSCAN struggled due to the smooth nature of the latent space, which does not create well-defined, density-separated clusters. As the VAE is designed for continuous and smooth latent spaces, it naturally results in more interpolation rather than discrete boundaries, which limits DBSCAN's effectiveness.

In comparison to the PCA baseline, VAE-based methods performed significantly better, capturing non-linear patterns in the data that PCA could not. The multi-modal VAE not only outperformed the audio-only models but also demonstrated the advantages of combining complementary data sources. By learning from both audio and text, the model could generate more informative representations, which led to better clustering results.

7 Limitations

Despite the promising results, several limitations should be considered:

Limited Dataset: The dataset used in this study, FMA-small, contains a limited number of tracks with lyrics, which restricts the effectiveness of multi-modal learning.

Metadata Quality: The fallback metadata (e.g., artist names, song titles) for songs without lyrics is less informative than the actual lyrics, limiting the feature quality for clustering.

Hyperparameter Sensitivity: The Beta-VAE model is sensitive to the choice of β. While $\beta = 4.0$ provided the best results, higher values caused over-regularization, leading to degraded performance.

Scalability: While the models performed well on the smaller dataset, scalability to larger datasets with more genres and longer tracks remains a concern.

Data Quality: Real-world music data often contains noise, missing values, and inconsistencies, which pose challenges for the robustness and generalization of the models.

8 Conclusion

This project demonstrated the effectiveness of Variational Autoencoders (VAEs) for unsupervised music clustering, particularly with the integration of multi-modal features, such as audio and text. The multi-modal VAE outperformed audio-only models by capturing richer representations, and the Beta-VAE model with $\beta = 4.0$ provided the best performance by learning disentangled latent variables, resulting in better cluster separability. Despite these improvements, challenges such as genre overlap, the use of single-genre labels, and limitations in metadata quality affected clustering results. The dataset's small size and reliance on fallback metadata for songs without lyrics also restricted the model's full potential.

Despite these limitations, the results highlight the promise of VAEs and multi-modal learning for music analysis. Future work should focus on expanding the dataset, improving metadata quality, and refining model architectures to handle the complexities of real-world music data. This approach offers significant potential for advancing unsupervised music clustering and other related applications in the field.

References

[1] Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

[2] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017.

[3] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. FMA: A dataset for music analysis. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pages 226–231, 1996.

[5] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[6] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[7] C.-C. Tseng, H.-C. Hsu, and C.-J. Lin, "Music Genre Classification Using Hybrid Feature Extraction and K-means Clustering," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 980-988, 2012.

[8] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[9] M. Bulat, M. Yang, and R. K. Grosse, "Audio Representation Learning with Deep Variational Autoencoders," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 85-99, 2019.

[10] C. Chu, Z. Shi, and L. Wang, "Music Genre Classification Using Variational Autoencoders," *Journal of Machine Learning Research*, vol. 20, no. 1, pp. 1-21, 2019.

[11] Y. Miao, I. King, and X. Liao, "Monotonic Multimodal VAE for Learning from Audio-Text Pairs," in *Proceedings of the 34th International Conference on Machine Learning*, 2016.

[12] I. Higgins, L. Matthey, A. Pal, et al., "Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[13] der Maaten, L. v., and Hinton, G., "Visualizing Data Using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579-2605, 2008.

[14] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *arXiv preprint arXiv:1802.03426*, 2018.