# KV Multimedia Search and Retrieval. Group C Report.

Maxim Kalpin
Johannes Kepler University Linz
k12140222@students.jku.at

Daniil Krechko
Johannes Kepler University Linz
k12149099@students.jku.at

Anastasia Ignatiuc
Johannes Kepler University Linz
k12137168@students.jku.at

Ilia Stepanov
Johannes Kepler University Linz
k12226374@students.jku.at

Artur Garipov
Johannes Kepler University Linz
k12146699@students.jku.at

## ABSTRACT

The exponential growth of multimedia data necessitates advanced search and retrieval systems capable of effectively handling diverse data modalities. This report presents the development of a multimodal music retrieval system utilizing the Music4All-Onion dataset. The system emphasizes content-driven retrieval by integrating textual, audio, and visual features to enhance retrieval accuracy and user experience. Through comprehensive experimental setups and evaluations, the proposed approach demonstrates superior performance compared to baseline retrieval systems in retrieving semantically relevant music tracks. Additionally, the system addresses beyond-accuracy metrics such as coverage, diversity, and popularity to ensure a balanced and user-centric retrieval process. The findings highlight the potential of multimodal integration in advancing music information retrieval systems.

**ACM Reference Format:**
Maxim Kalpin, Daniil Krechko, Anastasia Ignatiuc, Ilia Stepanov, and Artur Garipov. 2025. KV Multimedia Search and Retrieval. Group C Report.. In *Proceedings of (MMSR WS24/25)*. ACM, New York, NY, USA, 12 pages. https://doi.org/12.345/678_9

## 1 INTRODUCTION

In the era of digital information, the volume of multimedia data has surged exponentially, presenting both opportunities and challenges in the realm of information retrieval. Music, as a pervasive form of multimedia, not only serves as a primary medium for artistic expression but also as a significant component of digital libraries and streaming platforms. Effective retrieval of music information is paramount for enhancing user experience, facilitating discovery, and supporting various applications such as recommendation systems, music analysis, and content management.

Traditional music retrieval systems have predominantly relied on metadata-based approaches, utilizing information such as song titles, artists, genres, and user-generated tags to facilitate search and discovery. While these methods offer a foundational framework, they often fall short in capturing the nuanced and multifaceted nature of music content. The subjective perception of music, characterized by its auditory, lyrical, and visual elements, demands more sophisticated retrieval mechanisms that transcend mere textual metadata.

To address these limitations, recent advancements have focused on content-driven retrieval systems that leverage multimodal data integration. By incorporating diverse data modalities—such as audio features, lyrics, and visual representations—these systems aim to provide a more holistic and semantically rich retrieval experience. The integration of multimodal data not only enhances the accuracy of retrieval but also enriches the diversity and coverage of the results, aligning more closely with user intent and preferences.

This report delineates the development of a multimodal music retrieval system using the Music4All-Onion dataset. The primary objective of the project is to design and implement a system that amalgamates textual, audio, and visual features to enable robust and content-aware music retrieval. The system is structured to perform content-driven retrieval by evaluating and integrating various feature representations, thereby optimizing for both accuracy and beyond-accuracy metrics such as coverage, diversity, and popularity.

The methodology encompasses the creation of baseline retrieval systems, the incorporation of additional modalities beyond text-based features, and the exploration of fusion techniques to amalgamate different retrieval strategies. Experimental setups are meticulously designed to evaluate the performance of each retrieval system using standard accuracy metrics alongside novel beyond-accuracy metrics. The findings from these evaluations underscore the efficacy of multimodal integration in enhancing retrieval performance and offer insights into the trade-offs between different system objectives.

In summary, this report contributes to the field of music information retrieval by presenting a comprehensive approach to developing a multimodal retrieval system. The integration of diverse data modalities, coupled with a rigorous evaluation framework, demonstrates the potential of content-driven retrieval systems in delivering more accurate, diverse, and user-aligned music retrieval outcomes.

## 2 METHODOLOGY

The development of the multimodal music retrieval system is structured around a comprehensive methodology that encompasses data preprocessing, feature extraction and integration, the design of diverse retrieval algorithms, the application of fusion techniques, and a robust evaluation framework. This section details the systematic approaches employed to leverage the multifaceted data modalities within the Music4All-Onion dataset, aiming to facilitate effective content-driven music retrieval.

### 2.1 Data Preprocessing

Data preprocessing is a pivotal step to ensure the quality and consistency of the dataset, which directly impacts the performance of the retrieval system. The Music4All-Onion

*MMSR WS24/25, January 01–08, 2025, Linz, Austria*
2025. ACM ISBN 123-4567-89-101/11/21
https://doi.org/12.345/678_9

dataset comprises various data types, including textual metadata, audio features, visual features, and user-generated tags. The preprocessing pipeline involves data cleaning, merging, normalization, and filtering to construct a unified and reliable dataset.

*2.1.1 Data Cleaning and Merging.* The dataset is initially segmented into multiple TSV (Tab-Separated Values) files, each containing different aspects of the music tracks. The primary steps in data cleaning and merging include:

- **Parsing and Standardizing Genres**: The `genre` field, originally presented as a comma-separated string, is converted into a list format. This standardization facilitates easier manipulation and analysis. The top genre for each track is extracted to serve as the primary relevance criterion in retrieval evaluations.
- **Tag Extraction and Filtering**: User-generated tags are parsed from dictionary strings into list structures. Genre-specific tags are excluded to focus on descriptive and thematic tags. Additionally, tags with weights below a predefined threshold are filtered out to retain only significant tags.
- **Handling Missing Values**: Tracks with missing critical information, such as `popularity` scores, are identified and excluded to maintain data consistency. Specifically, tracks lacking `popularity` metrics are removed from the catalog to ensure the reliability of popularity-based evaluation metrics.

Subsequent merging operations integrate the cleaned datasets into a unified `catalog_df` DataFrame. This consolidation encompasses metadata, genres, tags, and all feature vectors, resulting in a comprehensive representation of each music track.

*2.1.2 Feature Normalization.* To ensure comparability across different feature types and scales, feature vectors undergo normalization using the L2 norm. This normalization step is crucial for similarity computations, particularly when employing distance-based metrics such as cosine similarity. Mathematically, the L2 normalization of a feature vector $\mathbf{f}$ is defined as:

$$\mathbf{f}_{\text{normalized}} = \frac{\mathbf{f}}{||\mathbf{f}||_2}$$

where $||\mathbf{f}||_2$ denotes the Euclidean norm of $\mathbf{f}$.

## 2.2 Feature Extraction and Integration

The retrieval system leverages multiple feature modalities extracted from the dataset, each contributing unique information to capture the multifaceted nature of music tracks. This section details the extraction and integration of textual, audio, visual, and tag features, providing a summary of the feature extraction processes carried out as part of the Music4All-Onion dataset development, detailing how various feature representations were derived from the raw data.

*2.2.1 Textual Features.*

- **BERT Embeddings**: Lyrics are processed using Bidirectional Encoder Representations from Transformers (BERT) to capture contextual semantic information. BERT embeddings provide dense vector representations that encapsulate the nuances of language within the lyrics, enabling the retrieval system to understand deeper semantic relationships.

*2.2.2 Audio Features.*

- **Mel-Frequency Cepstral Coefficients (MFCC)**: MFCCs are extracted to represent the short-term power spectrum of audio signals. These coefficients are widely used in audio processing for tasks such as speech and music recognition due to their effectiveness in capturing the timbral characteristics of audio.
- **Spectral Contrast**: This feature captures the difference in amplitude between peaks and valleys in the spectral envelope, providing insights into the timbral texture of the audio. Spectral contrast is instrumental in distinguishing between different musical genres and styles.

*2.2.3 Visual Features.*

- **VGG19 and ResNet Embeddings**: Visual features are extracted from YouTube videoclips associated with each track using deep convolutional neural networks, specifically VGG19 and ResNet architectures. These embeddings capture high-level visual patterns and aesthetics from the videoclips, which can be indicative of the music genre or mood, thereby enhancing the retrieval system's ability to associate visual cues with musical content. These embeddings capture high-level visual patterns and aesthetics, which can be indicative of the music genre or mood, thereby enhancing the retrieval system's ability to associate visual cues with musical content.

*2.2.4 Tag Features.*

- **Tag Vectorization**: User-generated tags are vectorized using both Term Frequency-Inverse Document Frequency (TF-IDF) and binary encoding. The TF-IDF vectorization quantifies the importance of tags within each song relative to the entire corpus, while binary encoding captures the presence or absence of tags. These vectorizations facilitate the computation of similarity based on semantic relevance derived from user-generated descriptors.

*2.2.5 Feature Integration.* Each feature modality is processed and stored as separate matrices to maintain modularity and facilitate independent analysis. For fusion techniques, feature matrices are combined either at the early stage (early fusion) by concatenating feature vectors or at the late stage (late fusion) by aggregating similarity scores derived from individual retrieval systems.

## 2.3 Retrieval Systems

The retrieval system comprises multiple retrieval approaches, each leveraging different feature modalities. These systems range from simple baselines to sophisticated multimodal retrieval methods, enabling a comprehensive evaluation of various retrieval strategies.

### 2.3.1 *Baseline Retrieval.*

*Random Retrieval.* The baseline system randomly selects $N$ tracks from the catalog, excluding the query track. This method serves as a reference point to evaluate the performance improvements introduced by more advanced retrieval techniques. Mathematically, for a query track $q$, the retrieved set $R(q)$ is:

$$R(q) = \text{RandomSample}(\text{Catalog} \setminus \{q\}, N)$$

where RandomSample denotes the random selection process.

### 2.3.2 *Text-Based Retrieval.*

*TF-IDF Retrieval.* Utilizing the TF-IDF representations of user-generated tags, this retrieval system computes cosine similarity between the query track and all other tracks to rank and retrieve the top $N$ similar tracks. The similarity score between two tracks $d_i$ and $d_j$ is defined as:

$$\text{Similarity}(d_i, d_j) = \frac{\mathbf{TFIDF}_{d_i} \cdot \mathbf{TFIDF}_{d_j}}{\|\mathbf{TFIDF}_{d_i}\|_2 \times \|\mathbf{TFIDF}_{d_j}\|_2}$$

where $\mathbf{TFIDF}_{d_i}$ and $\mathbf{TFIDF}_{d_j}$ are the TF-IDF vectors for tags of tracks $d_i$ and $d_j$, respectively.

### 2.3.3 *Multimodal Retrieval Systems.*

*BERT Retrieval.* This system employs BERT embeddings of lyrics to compute cosine similarity, capturing deeper semantic relationships compared to traditional TF-IDF approaches. The similarity score is computed as:

$$\text{Similarity}(d_i, d_j) = \frac{\mathbf{BERT}_{d_i} \cdot \mathbf{BERT}_{d_j}}{\|\mathbf{BERT}_{d_i}\|_2 \times \|\mathbf{BERT}_{d_j}\|_2}$$

where $\mathbf{BERT}_{d_i}$ and $\mathbf{BERT}_{d_j}$ are the BERT embedding vectors for tracks $d_i$ and $d_j$, respectively.

*MFCC Retrieval.* Using MFCC feature vectors, this system calculates Euclidean distances to identify the closest audio matches to the query track. The distance between two tracks $d_i$ and $d_j$ is given by:

$$\text{Distance}(d_i, d_j) = \|\mathbf{MFCC}_{d_i} - \mathbf{MFCC}_{d_j}\|_2$$

where $\mathbf{MFCC}_{d_i}$ and $\mathbf{MFCC}_{d_j}$ are the MFCC vectors for tracks $d_i$ and $d_j$, respectively.

*Spectral Contrast Retrieval.* Spectral contrast features are utilized to compute cosine similarity, highlighting tracks with similar timbral textures. The similarity score is defined as:

$$\text{Similarity}(d_i, d_j) = \frac{\mathbf{Spectral}_{d_i} \cdot \mathbf{Spectral}_{d_j}}{\|\mathbf{Spectral}_{d_i}\|_2 \times \|\mathbf{Spectral}_{d_j}\|_2}$$

where $\mathbf{Spectral}_{d_i}$ and $\mathbf{Spectral}_{d_j}$ are the spectral contrast vectors for tracks $d_i$ and $d_j$, respectively.

*VGG19 and ResNet Retrieval.* Visual embeddings from VGG19 and ResNet models are used to compute cosine similarity, retrieving tracks with visually similar album artwork. The similarity score is calculated as:

$$\text{Similarity}(d_i, d_j) = \frac{\mathbf{VGG19}_{d_i} \cdot \mathbf{VGG19}_{d_j}}{\|\mathbf{VGG19}_{d_i}\|_2 \times \|\mathbf{VGG19}_{d_j}\|_2}$$

$$\text{Similarity}(d_i, d_j) = \frac{\mathbf{ResNet}_{d_i} \cdot \mathbf{ResNet}_{d_j}}{\|\mathbf{ResNet}_{d_i}\|_2 \times \|\mathbf{ResNet}_{d_j}\|_2}$$

where $\mathbf{VGG19}_{d_i}$ and $\mathbf{VGG19}_{d_j}$ are the VGG19 feature vectors, and $\mathbf{ResNet}_{d_i}$ and $\mathbf{ResNet}_{d_j}$ are the ResNet feature vectors for tracks $d_i$ and $d_j$, respectively.

*Tag-Based Retrieval.* This system leverages the binary tag matrix to compute cosine similarity based on user-generated tags, emphasizing semantic relevance derived from descriptive tags. The similarity score is computed as:

$$\text{Similarity}(d_i, d_j) = \frac{\mathbf{Tag}_{d_i} \cdot \mathbf{Tag}_{d_j}}{\|\mathbf{Tag}_{d_i}\|_2 \times \|\mathbf{Tag}_{d_j}\|_2}$$

where $\mathbf{Tag}_{d_i}$ and $\mathbf{Tag}_{d_j}$ are the binary tag vectors for tracks $d_i$ and $d_j$, respectively.

## 2.4 Fusion Techniques

To harness the strengths of individual retrieval systems, fusion techniques are employed to integrate multiple modalities, thereby enhancing overall retrieval performance. The two primary fusion strategies utilized are early fusion and late fusion.

2.4.1 *Early Fusion.* Early fusion involves concatenating feature vectors from different modalities into a single, unified feature vector before similarity computation. In our approach, BERT embeddings and MFCC features are combined to form a comprehensive feature representation for each track. Mathematically, the combined feature vector $\mathbf{f}_{\text{combined}}$ for a track is:

$$\mathbf{f}_{\text{combined}} = [\mathbf{f}_{\text{BERT}}; \mathbf{f}_{\text{MFCC}}]$$

where $\mathbf{f}_{\text{BERT}}$ and $\mathbf{f}_{\text{MFCC}}$ are the BERT and MFCC feature vectors, respectively. Cosine similarity is then computed on these combined vectors to retrieve similar tracks, leveraging the complementary information from both textual and audio modalities.

2.4.2 *Late Fusion.* Late fusion aggregates similarity scores from different retrieval systems to compute an overall similarity score. Specifically, the system combines similarity scores from the MFCC Retrieval and VGG19 Retrieval systems using a weighted sum. Given two similarity scores $s_1$ and $s_2$ from MFCC and VGG19 retrievals, respectively, the aggregated similarity $s_{\text{agg}}$ is:

$$s_{\text{agg}} = \alpha s_1 + \beta s_2$$

where $\alpha$ and $\beta$ are weights assigned to each modality, reflecting their relative importance. In our implementation, $\alpha$ is set to 0.5, and $\beta = 1 - \alpha$. The final ranking is based on the aggregated similarity scores, allowing the system to balance the contributions of audio and visual modalities effectively.

## 2.5 Retrieval Algorithms

Each retrieval algorithm is encapsulated within a dedicated function, adhering to a consistent interface that accepts a query track ID and returns a ranked list of similar tracks based on the respective similarity measures. This modular design facilitates seamless integration and evaluation within the overall system architecture.

*2.5.1 Random Retrieval.* As described in the baseline retrieval, the Random Retrieval algorithm serves as a benchmark, providing a reference point for evaluating the efficacy of more sophisticated retrieval techniques. Since this method selects tracks randomly, it does not utilize any similarity measure.

*2.5.2 TF-IDF Retrieval.* Leveraging TF-IDF representations of user-generated tags, this algorithm computes cosine similarity to rank tracks based on semantic relevance derived from tag information.

*Rationale for Cosine Similarity.* Cosine similarity is chosen for TF-IDF Retrieval because it effectively measures the orientation between high-dimensional sparse vectors, which is characteristic of TF-IDF tag representations. Since TF-IDF vectors are normalized and sparse, cosine similarity is well-suited to capture the semantic similarity between tracks based on their tag distributions, regardless of the magnitude of the vectors. This focus on angular similarity ensures that tracks with similar tag patterns are ranked closely, enhancing the retrieval of semantically relevant tracks.

*2.5.3 BERT Retrieval.* Utilizing BERT embeddings, this algorithm captures deeper semantic relationships within the lyrics, enhancing the retrieval of contextually relevant tracks through cosine similarity.

*Rationale for Cosine Similarity.* BERT embeddings produce dense, high-dimensional vectors that encapsulate rich semantic information from lyrics. Cosine similarity is ideal for these embeddings as it measures the cosine of the angle between two vectors, effectively assessing their directional alignment in the embedding space. This measure is particularly effective for dense representations where the magnitude of the vectors is less important than their direction, allowing the system to identify tracks with similar lyrical semantics.

*2.5.4 MFCC Retrieval.* By calculating Euclidean distances between MFCC feature vectors, this algorithm identifies tracks with similar audio characteristics, emphasizing timbral and rhythmic similarities.

*Rationale for Euclidean Distance.* Mel-Frequency Cepstral Coefficients (MFCC) are designed to capture the timbral aspects of audio signals, resulting in feature vectors that represent the short-term power spectrum of sound. Euclidean distance is chosen for MFCC Retrieval because it quantifies the absolute differences between feature vectors, making it effective for capturing the precise acoustic similarities and dissimilarities between tracks. Since MFCC vectors are typically dense and represent continuous audio features, Euclidean distance provides a straightforward and intuitive measure of similarity based on the actual feature values.

*2.5.5 Spectral Contrast Retrieval.* This algorithm computes cosine similarity based on spectral contrast features, targeting the retrieval of tracks with similar timbral textures.

*Rationale for Cosine Similarity.* Spectral contrast features capture the differences in amplitude between peaks and valleys in the spectral envelope, providing insights into the timbral texture of audio. These features result in high-dimensional vectors where the direction of the vector is more indicative of similarity than its magnitude. Cosine similarity is thus appropriate as it assesses the angular similarity between spectral contrast vectors, ensuring that tracks with similar timbral textures are identified and ranked closely, irrespective of the overall energy levels.

*2.5.6 VGG19 and ResNet Retrieval.* Using visual embeddings from VGG19 and ResNet models, these algorithms retrieve tracks with visually similar YouTube videoclips, incorporating aesthetic dimensions into the retrieval process through cosine similarity.

*Rationale for Cosine Similarity.* Visual embeddings extracted from deep convolutional neural networks like VGG19 and ResNet produce dense, high-dimensional vectors that encode complex visual patterns and aesthetics from YouTube videoclips. Cosine similarity is chosen because it effectively measures the similarity in the direction of these dense vectors, making it suitable for capturing nuanced visual similarities without being affected by the magnitude of the embeddings. This ensures that tracks with visually similar content are accurately identified, enhancing the multimodal retrieval capabilities.

*2.5.7 Tag-Based Retrieval.* Employing the binary tag matrix, this algorithm computes cosine similarity based on user-generated tags, emphasizing semantic relevance derived from descriptive tags.

*Rationale for Cosine Similarity.* The binary tag matrix represents the presence or absence of specific tags for each track, resulting in high-dimensional sparse vectors. Cosine similarity is ideal for such representations as it effectively measures the angular similarity between sparse vectors, focusing on the overlap of tags irrespective of the vector magnitudes. This makes it well-suited for identifying tracks that share similar descriptive tags, ensuring that semantically relevant tracks are retrieved.

## 2.6 Fusion Techniques

To harness the strengths of individual retrieval systems, fusion techniques are employed to integrate multiple modalities, thereby enhancing overall retrieval performance.

*2.6.1 Early Fusion.* Early fusion involves concatenating feature vectors from different modalities into a single, unified feature vector before similarity computation. In our approach, BERT embeddings and MFCC features are combined to form a comprehensive feature representation for each track. The combined feature vector $\mathbf{f}_{\text{combined}}$ is then used to compute cosine similarity, facilitating the retrieval of tracks that are semantically and audibly similar to the query.

*2.6.2 Late Fusion.* Late fusion aggregates similarity scores from different retrieval systems to compute an overall similarity score. Specifically, the system combines similarity scores from the MFCC Retrieval and VGG19 Retrieval systems using a weighted sum. Given two similarity scores $s_1$ and $s_2$ from MFCC and VGG19 retrievals, respectively, the aggregated similarity $s_{\text{agg}}$ is:

$$s_{\text{agg}} = \alpha s_1 + \beta s_2$$

where $\alpha$ and $\beta$ are weights assigned to each modality, reflecting their relative importance. In our implementation, $\alpha$ is set to 0.5, and $\beta = 1 - \alpha$. These weights were initially chosen to balance the contributions of audio and visual modalities equally. Subsequent parameter tuning was conducted using a grid search approach, evaluating combinations of $\alpha$ and $\beta$ to maximize retrieval performance based on Precision@10 and NDCG@10 metrics. The selected weights demonstrated the best performance, ensuring an optimal balance between the modalities.

## 2.7 Evaluation Metrics

The retrieval systems are evaluated using a combination of accuracy and beyond-accuracy metrics to assess both the relevance and quality of the retrieved results. This dual-faceted evaluation ensures a comprehensive understanding of system performance from multiple perspectives. All metrics, except for Coverage@10, are computed for each query individually and then averaged across all queries to obtain a comprehensive performance overview.

*2.7.1 Accuracy Metrics.*

- **Precision@$k$**: Measures the proportion of retrieved tracks that are relevant.

$$\text{Precision@}k = \frac{|\text{Retrieved} \cap \text{Relevant}|}{k}$$

- **Recall@$k$**: Measures the proportion of relevant tracks that are retrieved.

$$\text{Recall@}k = \frac{|\text{Retrieved} \cap \text{Relevant}|}{|\text{Relevant}|}$$

- **Normalized Discounted Cumulative Gain (NDCG@$k$)**: Evaluates the ranking quality by considering the position of relevant tracks.

$$\text{NDCG@}k = \frac{\text{DCG@}k}{\text{IDCG@}k}$$

where DCG@$k$ is the Discounted Cumulative Gain and IDCG@$k$ is the Ideal DCG.

- **Mean Reciprocal Rank (MRR)**: Calculates the average of the reciprocal ranks of the first relevant track in the retrieved list.

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$$

*2.7.2 Beyond-Accuracy Metrics.*

- **Coverage@$N$**: Represents the percentage of tracks that appear in at least one retrieval list across all queries.

$$\text{Coverage@}N = \left( \frac{|\bigcup_{q \in Q} \text{Retrieved}(q)|}{|\text{Catalog}|} \right) \times 100\%$$

- **Tag Diversity@$N$**: Assesses the average number of unique tags among the top $N$ retrieved tracks.

$$\text{Tag Diversity@}N = \frac{1}{|Q|} \sum_{q \in Q} |\text{Unique Tags}(\text{Retrieved}(q))|$$

- **Popularity Diversity@$N$**: Measures the variance in popularity scores among the top $N$ retrieved tracks. The `popularity` score for each track is a numerical value representing its overall popularity on the platform, calculated based on factors such as play counts, user ratings, and engagement metrics. A higher variance indicates a broader range of popularity levels among the retrieved tracks.

$$\text{Popularity Diversity@}N = \frac{1}{|Q|} \sum_{q \in Q} \text{Var}(\text{Popularity}(\text{Retrieved}(q)))$$

- **Average Popularity@$N$**: Computes the mean popularity score of the top $N$ retrieved tracks. This metric provides an indication of the overall popularity of the retrieved set, with higher values signifying that, on average, more popular tracks are being retrieved.

$$\text{AvgPop@}N = \frac{1}{|Q| \times N} \sum_{q \in Q} \sum_{t \in \text{Retrieved}(q)} \text{Popularity}(t)$$

*2.7.3 Trade-offs and Optimization.* The evaluation framework not only measures the accuracy and quality of retrievals but also explores the trade-offs between different metrics. Specifically, the system investigates how optimizing for diversity metrics such as Tag Diversity and Popularity Diversity may impact accuracy metrics like NDCG@10. Parameter tuning and controlled experiments are conducted to balance these objectives, ensuring that the retrieval system delivers both relevant and diverse results.

## 2.8 Implementation Details

The retrieval system was implemented using Python, leveraging a suite of libraries to facilitate data manipulation, feature extraction, similarity computations, and evaluation. Key libraries and tools include:

- **Pandas**: For efficient data manipulation and preprocessing.
- **Scikit-learn**: Utilized for feature extraction (TF-IDF, CountVectorizer), normalization, and similarity computations (cosine similarity).
- **SciPy**: Employed for handling sparse matrices and performing efficient numerical operations.
- **NLTK**: Used for natural language processing tasks, including tokenization and lemmatization.
- **Matplotlib and Seaborn**: For data visualization and exploratory analysis.

#### 2.8.1 System Architecture.

The system architecture is modular, comprising distinct components for data preprocessing, feature extraction, retrieval algorithms, fusion techniques, and evaluation. This modularity facilitates scalability, allowing for the seamless addition of new feature modalities and retrieval strategies.

#### 2.8.2 Retrieval Algorithms.

Each retrieval algorithm is encapsulated within a dedicated function, adhering to a consistent interface that accepts a query track ID and returns a ranked list of similar tracks based on the respective similarity measures. This design ensures interoperability and simplifies the evaluation process across different retrieval systems.

#### 2.8.3 Fusion Methods.

Early and late fusion techniques are implemented as separate modules, enabling independent experimentation and optimization. The early fusion module concatenates BERT and MFCC feature vectors before similarity computation, while the late fusion module aggregates similarity scores from the MFCC Retrieval and VGG19 Retrieval systems using weighted combinations.

#### 2.8.4 Reproducibility and Source Code Management.

To ensure reproducibility and facilitate collaborative development, the source code is hosted on GitHub[1]. The repository contains well-documented scripts, organized directories for data and results, and comprehensive instructions for setting up the environment and running experiments. Additionally, version control practices are adhered to, maintaining a clear history of changes and updates to the codebase.

### 2.9 Experimental Setup

The experimental evaluation encompasses the following components to ensure a thorough assessment of the retrieval systems:

- **Dataset**: The Music4All-Onion dataset, comprising 5148 music tracks with comprehensive feature sets across textual, audio, visual, and tag modalities, serves as the foundation for all experiments.
- **Queries**: Each track in the dataset functions as a query, facilitating an exhaustive evaluation across all entries and ensuring robustness in performance assessments.
- **Retrieval Configurations**: Ten retrieval systems are evaluated, including baseline, text-based, individual modality-based, and fusion-based approaches. This diverse set of configurations allows for comparative analyses and the identification of optimal retrieval strategies.
- **Evaluation Metrics**: A combination of accuracy and beyond-accuracy metrics is employed to assess the performance comprehensively. This dual-metric approach ensures that the retrieval systems are evaluated not only for relevance but also for quality aspects such as diversity and coverage.
- **Reproducibility**: To ensure that the experiments can be replicated, all random processes are seeded appropriately, and detailed documentation is provided within the codebase. Additionally, intermediate results and logs are systematically stored for verification and analysis purposes.

#### 2.9.1 Experimental Procedure.

The experimental procedure follows a structured workflow:

1. **Data Preparation**: Load and preprocess the dataset, including data cleaning, merging, and normalization.
2. **Feature Extraction**: Extract and normalize features across all modalities—textual, audio, visual, and tags.
3. **Retrieval System Implementation**: Implement the baseline and advanced retrieval algorithms, ensuring adherence to a consistent interface for seamless evaluation.
4. **Fusion Techniques Application**: Apply early and late fusion techniques to integrate multiple modalities, enhancing the retrieval performance.
5. **Evaluation**: Execute the retrieval systems against all queries, computing the defined accuracy and beyond-accuracy metrics to assess performance comprehensively.
6. **Analysis and Discussion**: Analyze the evaluation results to identify strengths and weaknesses of each retrieval system, discussing the trade-offs between different metrics and the impact of multimodal integration.

#### 2.9.2 Parameter Tuning and Optimization.

To optimize the retrieval systems, parameter tuning is conducted iteratively. Parameters such as the number of features in vector representations, weighting factors in fusion techniques, and threshold values for tag filtering are adjusted based on preliminary evaluation results. Grid search and cross-validation techniques are employed to identify the optimal parameter configurations that balance accuracy and beyond-accuracy objectives effectively.

#### 2.9.3 Relevance Definitions.

The evaluation framework considers multiple relevance definitions to capture different aspects of track similarity:

- **Top Genre Matching**: A retrieved track is deemed relevant if its top genre matches the top genre of the query track. This definition emphasizes genre-based similarity.
- **Tag Overlap**: A retrieved track is considered relevant if it shares a significant number of tags with the query track, capturing thematic and descriptive similarities beyond genre.

By evaluating the retrieval systems under different relevance definitions, the system's robustness and versatility are thoroughly assessed.

## 3 RESULTS AND DISCUSSION

This section presents the evaluation outcomes of the developed multimodal music retrieval systems. The results are categorized into quantitative metrics and qualitative analyses, providing a comprehensive overview of each system's performance and capabilities.

[1] https://github.com/Goldenwert/mmsr_ws24_c

## 3.1 Quantitative Evaluation

The retrieval systems were assessed using a suite of accuracy and beyond-accuracy metrics. Comprehensive results for all retrieval systems are provided in Appendix B.

*3.1.1 Top Genre Relevance.* Under the *top_genre* relevance definition, a retrieved track is considered relevant if its top genre matches that of the query track. This definition emphasizes genre-based similarity, aligning retrieval performance with genre consistency.

- **Precision@10**: Among all retrieval systems, **Late Fusion MFCC+VGG19 Retrieval** achieved the highest Precision@10 (0.099), closely followed by **Early Fusion BERT+MFCC Retrieval** (0.098) and **Tag-Based Retrieval** (0.095). These systems significantly outperform the baseline **Random Retrieval** (0.042), indicating the effectiveness of fusion-based and tag-based approaches in identifying relevant tracks based on genre consistency.
- **Recall@10**: **Late Fusion MFCC+VGG19 Retrieval**, **Early Fusion BERT+MFCC Retrieval**, and **MFCC Retrieval** each achieved a Recall@10 of 0.008. While these values are higher than those of **Random Retrieval** (0.002), they suggest that these systems have limited coverage of relevant tracks within the top 10 results.
- **NDCG@10**: **Late Fusion MFCC+VGG19 Retrieval** leads in NDCG@10 (0.108), followed by **Early Fusion BERT+MFCC Retrieval** and **MFCC Retrieval** (both 0.103). This indicates a superior ability to rank more relevant tracks higher in the retrieval list.
- **Mean Reciprocal Rank (MRR)**: The highest MRR is observed in **Late Fusion MFCC+VGG19 Retrieval** (0.231), indicating that the first relevant track retrieved by this system appears very early in the ranked list. **Early Fusion BERT+MFCC Retrieval** and **ResNet Retrieval** also demonstrate strong performance with MRR values of 0.208.

*3.1.2 Tag Overlap Relevance.* Under the *tag_overlap* relevance definition, a retrieved track is considered relevant if it shares a significant number of tags with the query track. This definition emphasizes thematic and descriptive similarities beyond genre consistency.

- **Precision@10**: **Late Fusion MFCC+VGG19 Retrieval** achieves the highest Precision@10 (0.0025), followed by **ResNet Retrieval** (0.0018) and **VGG19 Retrieval** (0.0018). These systems outperform the baseline **Random Retrieval** (0.0002), albeit the improvement is marginal, highlighting the challenge of accurately identifying relevant tracks based solely on tag overlap.
- **Recall@10**: **Late Fusion MFCC+VGG19 Retrieval** also leads in Recall@10 (0.0060), surpassing all other systems, including **ResNet Retrieval** (0.0046) and **VGG19 Retrieval** (0.0047). This indicates a better coverage of relevant tracks within the top 10 results.
- **NDCG@10**: The highest NDCG@10 is observed in **Late Fusion MFCC+VGG19 Retrieval** (0.0066), followed by **ResNet Retrieval** (0.0056) and **VGG19 Retrieval** (0.0053). These values, although low, reflect the systems' ability to rank relevant tracks higher within the retrieval list.
- **Mean Reciprocal Rank (MRR)**: **Late Fusion Retrieval** achieves the highest MRR (0.0133), indicating that the first relevant track retrieved by this system appears earlier in the ranked list compared to others. **ResNet Retrieval** and **VGG19 Retrieval** follow with MRR values of 0.0119 and 0.0108, respectively.

*3.1.3 Beyond-Accuracy Metrics.* Beyond-accuracy metrics such as Coverage@10, Tag Diversity@10, Popularity Diversity@10, and AvgPop@10 provide a broader perspective on the retrieval system's performance, assessing aspects like the variety and popularity of retrieved tracks.

- **Coverage@10**: **Random Retrieval** achieves full coverage (100.00%), meaning every track appears in at least one retrieval list. **Tag-Based Retrieval** has a coverage of 88.17%, indicating that some tracks are not retrieved by this system. **Early Fusion BERT + MFCC Retrieval** and **Late Fusion MFCC + VGG19 Retrieval** maintain moderate coverage at 91.06% and 83.60%, respectively.
- **Tag Diversity@10**: Most systems maintain high Tag Diversity@10 scores, ranging from 9.52 to 9.73. **Tag-Based Retrieval** exhibits a Tag Diversity@10 of 9.71, slightly lower than some fusion-based systems but still indicative of a wide range of tags among the retrieved tracks.
- **Popularity Diversity@10**: There is noticeable variability in Popularity Diversity@10 across systems. **Random Retrieval** has the highest Popularity Diversity@10 (191.58), reflecting a broad range of popularity levels in its selections. **Tag-Based Retrieval** achieves a moderate Popularity Diversity@10 (161.03), while fusion-based systems like **Late Fusion MFCC + VGG19 Retrieval** maintain a balanced Popularity Diversity@10 (176.08).
- **AvgPop@10**: **Late Fusion MFCC + VGG19 Retrieval** achieves the highest AvgPop@10 (37.84), indicating that it retrieves more popular tracks on average. **Tag-Based Retrieval** balances relevance and popularity with an AvgPop@10 of 34.99, while **Random Retrieval** maintains an AvgPop@10 of 35.11.

Overall, **Late Fusion MFCC + VGG19 Retrieval** consistently outperforms other systems across all accuracy metrics, highlighting the effectiveness of leveraging both MFCC and VGG19 features in a fusion-based approach for music retrieval. **Tag-Based Retrieval** offers strong precision and recall but with slightly lower diversity, and fusion-based approaches demonstrate the advantage of combining multiple modalities for improved retrieval performance.

## 3.2 Qualitative Analysis

While quantitative metrics provide a numerical assessment of retrieval performance, qualitative analysis offers deeper insights into the system's behavior and the nature of the retrieved results. This subsection explores a specific query example, examining how different retrieval systems respond to the query "*Laura Pausini - Tra Te E Il Mare*."

### 3.2.1 *Example Query: "Laura Pausini - Tra Te E Il Mare".*

To provide deeper insights into the retrieval systems' performance, a qualitative analysis was conducted using the query song "*Laura Pausini - Tra Te E Il Mare*". The top five retrieved tracks from each of the leading retrieval systems under the *top_genre* relevance definition are discussed below. Detailed lists of all retrieved tracks are available in Appendix A.

*Top Genre Relevance.* Under the *top_genre* relevance definition, the query "*Laura Pausini - Tra Te E Il Mare*" was evaluated to retrieve tracks with matching top genres. The following analysis highlights the performance and characteristics of each retrieval system based on the retrieved results.

- **Late Fusion MFCC+VGG19 Retrieval**: This system retrieved highly relevant tracks such as "Sheryl Crow - Superstar" (0.921) and "James Morrison - One Last Chance" (0.916). The high Precision@10 and NDCG@10 scores indicate effective genre-based matching and ranking.
- **Early Fusion BERT+MFCC Retrieval**: Retrieved tracks like "Britney Spears - Break the Ice" (0.942) and "Kylie Minogue - My Secret Heart" (0.939) demonstrate strong genre consistency, as reflected in its high Precision@10 and AvgPop@10 scores.
- **Tag-Based Retrieval**: Although primarily tag-oriented, this system successfully identified genre-consistent tracks such as "Billie Myers - Kiss The Rain" (0.617). However, the lower Tag Diversity@10 suggests that retrieved tracks may share similar tags.
- **MFCC Retrieval**: With tracks like "Britney Spears - Break the Ice" (9.916) and "Kylie Minogue - My Secret Heart" (10.493), this system emphasizes acoustic features that align well with the query's genre.
- **ResNet Retrieval**: Retrieved tracks such as "Miguel Bosé - Duende" (40.754) and "Bat for Lashes - A Wall" (41.078) showcase the system's ability to identify genre-consistent tracks based on visual embeddings.

### 3.2.2 *Analysis of "Laura Pausini - Tra Te E Il Mare" Query Results.*

The query "*Laura Pausini - Tra Te E Il Mare*," classified under its respective genres and tags, was evaluated across different retrieval systems. The following analysis highlights the performance and characteristics of each system based on the retrieved results.

*Late Fusion MFCC+VGG19 Retrieval.*

- **Top Genre Relevance**: Retrieved tracks like "Sheryl Crow - Superstar" (0.921) and "James Morrison - One Last Chance" (0.916) demonstrate high genre alignment. The system's fusion approach effectively combines acoustic and visual features, resulting in precise genre-based retrievals.

*Early Fusion BERT+MFCC Retrieval.*

- **Top Genre Relevance**: Retrieved tracks like "Britney Spears - Break the Ice" (0.942) and "Kylie Minogue - My Secret Heart" (0.939) showcase effective genre-based retrieval. The integration of BERT's contextual embeddings with MFCC features enhances the system's ability to capture both semantic and acoustic similarities.

*Tag-Based Retrieval.*

- **Top Genre Relevance**: Retrieved track "Billie Myers - Kiss The Rain" (0.617) highlights the system's ability to leverage tag-based similarity for genre-consistent retrieval. However, the lower Precision@10 compared to fusion-based systems suggests room for improvement in accurately identifying top-relevant tracks.

*MFCC Retrieval.*

- **Top Genre Relevance**: Tracks such as "Britney Spears - Break the Ice" (9.916) and "Kylie Minogue - My Secret Heart" (10.493) indicate strong acoustic feature alignment with the query's genre. The high Similarity Scores reflect the system's proficiency in identifying musically similar tracks.

*ResNet Retrieval.*

- **Top Genre Relevance**: Retrieved tracks such as "Miguel Bosé - Duende" (40.754) and "Bat for Lashes - A Wall" (41.078) demonstrate effective genre-based retrieval through visual embeddings. The system excels in identifying tracks that not only match the genre but also share visual aesthetic similarities.

### 3.2.3 *Discussion of Qualitative Findings.*

The qualitative analysis of the "*Laura Pausini - Tra Te E Il Mare*" query reveals the distinct operational characteristics of each retrieval system under the *top_genre* relevance definition:

- **Late Fusion MFCC+VGG19 Retrieval**:
  - **Strengths**: Excels in both precision and recall. The fusion of MFCC (acoustic features) and VGG19 (visual features) enables the system to effectively capture genre consistency, resulting in highly relevant and accurately ranked retrievals.
  - **Limitations**: While achieving high relevance, the system may retrieve tracks with varying popularity levels, as indicated by its high Popularity Diversity@10 (176.08).
- **Early Fusion BERT+MFCC Retrieval**:
  - **Strengths**: Effectively captures both semantic and acoustic similarities, leading to highly relevant retrievals. The integration of BERT's contextual embeddings with MFCC features enhances the system's ability to align retrieved tracks with the query's genre.
  - **Limitations**: Although precision is high, the system may retrieve tracks with moderate popularity diversity, indicating a balance between relevance and diversity.
- **Tag-Based Retrieval**:
  - **Strengths**: Demonstrates strong genre consistency through tag-based similarity, effectively retrieving tracks that align with the query's genre.
  - **Limitations**: The significantly lower Tag Diversity@10 (5.20) suggests that retrieved tracks may

share similar tags, potentially limiting the variety of genres or sub-genres represented.

- **MFCC Retrieval**:
  - **Strengths**: High Precision@10 indicates effective acoustic feature matching, ensuring that retrieved tracks are musically similar and genre-consistent.
  - **Limitations**: While acoustic similarity ensures genre consistency, the system may overlook thematic or lyrical similarities, limiting the scope of retrieved tracks.
- **ResNet Retrieval**:
  - **Strengths**: Excels in identifying genre-consistent tracks through visual embeddings, effectively capturing aesthetic similarities reflected in album artwork. This dual consideration of acoustic and visual features enhances the system's retrieval accuracy.
  - **Limitations**: The focus on visual similarity may lead to the retrieval of tracks that are visually similar but may not align perfectly with the genre nuances of the query.

## 3.3 Discussion

The evaluation results underscore the significance of multimodal integration in enhancing music retrieval performance. Fusion techniques, particularly the **Late Fusion MFCC+VGG19 Retrieval**, consistently outperform unimodal systems across both relevance definitions, highlighting the complementary strengths of audio and visual features in capturing both genre-based and thematic similarities.

*3.3.1 Impact of Fusion Techniques.* The superior performance of the late fusion approach suggests that aggregating similarity scores from distinct modalities can effectively balance the contributions of each feature type, leading to more robust and relevant retrievals. The integration of MFCC audio features with VGG19 visual embeddings allows the system to capture a broader spectrum of similarities, encompassing both auditory and visual aspects of music tracks.

In contrast, early fusion, while also effective, may face challenges related to feature vector dimensionality and the relative weighting of different modalities. The late fusion approach, by maintaining separate similarity computations before aggregation, provides greater flexibility in balancing modality contributions, thereby enhancing overall retrieval quality.

*3.3.2 Role of Individual Modalities.* Among unimodal systems, audio-based retrievals (**MFCC Retrieval**) demonstrate strong performance under the *top_genre* relevance definition, emphasizing the importance of audio features in capturing genre-specific characteristics. Visual-based retrievals (**VGG19 Retrieval**) also show commendable performance under the *tag_overlap* relevance definition, indicating their effectiveness in identifying thematically similar tracks based on visual aesthetics.

Textual features, represented through **BERT Retrieval** and **TF-IDF Retrieval**, exhibit moderate performance, suggesting that while they contribute valuable semantic information, their standalone effectiveness may be limited without integration with other modalities.

*3.3.3 Beyond-Accuracy Metrics.* The beyond-accuracy metrics, including Coverage@10, Tag Diversity@10, Popularity Diversity@10, and AvgPop@10, provide additional insights into the retrieval systems' performance. High Coverage@10 values across most systems indicate that the majority of tracks are retrievable, ensuring a comprehensive coverage of the catalog. Tag Diversity@10 and Popularity Diversity@10 metrics reveal that the retrieval systems maintain a balanced diversity in both descriptive tags and popularity scores, enhancing the user experience by providing varied and engaging results.

*3.3.4 Trade-offs and Optimization.* The evaluation highlights inherent trade-offs between different metrics. For instance, optimizing for higher Precision@10 may lead to reduced Diversity@10, as more similar tracks are retrieved at the expense of diversity. The fusion techniques offer a balanced approach, achieving high relevance while maintaining substantial diversity and coverage, thereby addressing these trade-offs effectively.

# 4 CONCLUSION

The evaluation of the multimodal music retrieval systems demonstrates the efficacy of integrating diverse data modalities to enhance retrieval performance. Fusion techniques, particularly late fusion, significantly improve relevance and diversity, outperforming unimodal approaches. The incorporation of audio and visual features, alongside textual information, provides a comprehensive framework for capturing the multifaceted nature of music tracks, leading to more relevant and engaging retrieval results.

Overall, the findings highlight the potential of multimodal integration in advancing music information retrieval systems, offering a balanced and user-centric approach to music discovery.

# A RETRIEVED TRACKS FOR QUERY: *LAURA PAUSINI - TRA TE E IL MARE*

The following ten tables present the top ten retrieved tracks for each retrieval system based on the query *Laura Pausini - Tra Te E Il Mare* under the *top_genre* relevance definition.

## A.1 Random Retrieval

**Table 1: Top 10 Retrieved Tracks using Random Retrieval under Top Genre Relevance**

| Artist | Song | Similarity Score |
|---|---|---|
| Jimmy Cliff | I Can See Clearly Now | 0.042 |
| 3OH!3 | House Party | 0.042 |
| Billie Myers | Kiss The Rain | 0.042 |
| Muddy Waters | I Just Want To Make Love To You | 0.042 |
| Forfun | Cósmica | 0.042 |
| Lenka | Everything at Once | 0.042 |
| Deafheaven | Worthless Animal | 0.042 |
| Wax | Right Between the Eyes | 0.042 |
| Robert Palmer | Looking For Clues | 0.042 |
| Tokio Hotel | Boy Don't Cry | 0.042 |

## A.2 Tag-Based Retrieval

**Table 2: Top 10 Retrieved Tracks using Tag-Based Retrieval under Top Genre Relevance**

| Artist | Song | Similarity Score |
|---|---|---|
| Billie Myers | Kiss The Rain | 0.617 |
| Hooverphonic | Frosted Flake Wood | 0.567 |
| Laura Pausini | La prospettiva di me | 0.567 |
| Aretha Franklin | What A Friend We Have In Jesus | 0.535 |
| 10,000 Maniacs | More Than This | 0.535 |
| L7 | Monster | 0.535 |
| Ambar Lucid | Eyes | 0.535 |
| Georgi Kay | Lone Wolf | 0.534 |
| Céline Dion | Coulda Woulda Shoulda | 0.534 |
| Missy Elliott | Bite Our Style (Interlude) | 0.534 |

## A.3 TF-IDF Retrieval

**Table 3: Top 10 Retrieved Tracks using TF-IDF Retrieval under Top Genre Relevance**

| Artist | Song | Similarity Score |
|---|---|---|
| Billie Myers | Kiss The Rain | 0.346 |
| Georgi Kay | Lone Wolf | 0.304 |
| Hooverphonic | Frosted Flake Wood | 0.244 |
| Aretha Franklin | What A Friend We Have In Jesus | 0.232 |
| Missy Elliott | Bite Our Style (Interlude) | 0.232 |
| Céline Dion | Coulda Woulda Shoulda | 0.232 |
| India.Arie | Good Man | 0.232 |
| Ambar Lucid | Eyes | 0.232 |
| Christina Perri | Shot Me in the Heart | 0.232 |
| L7 | Monster | 0.232 |

## A.4 BERT Retrieval

**Table 4: Top 10 Retrieved Tracks using BERT Retrieval under Top Genre Relevance**

| Artist | Song | Similarity Score |
|---|---|---|
| Avenged Sevenfold | So Far Away | 0.810 |
| Mon Laferte | Vuelve Por Favor | 0.802 |
| Ricky Martin | Vuelve | 0.788 |
| The Kooks | All the Time | 0.770 |
| La Oreja de Van Gogh | Flores en la orilla | 0.764 |
| Los Hermanos | Condicional | 0.763 |
| Staind | Home | 0.761 |
| Evanescence | My Immortal | 0.758 |
| Katie Melua | I Cried for You | 0.757 |
| Mary J. Blige | Come to Me (Peace) | 0.756 |

## A.5 MFCC Retrieval

**Table 5: Top 10 Retrieved Tracks using MFCC Retrieval under Top Genre Relevance**

| Artist | Song | Similarity Score |
|---|---|---|
| Britney Spears | Break the Ice | 9.916 |
| Kylie Minogue | My Secret Heart | 10.493 |
| Backstreet Boys | Safest Place to Hide | 10.517 |
| Foxy Shazam | Introducing Foxy | 10.928 |
| Skank | Esquecimento | 11.003 |
| Take That | Hold On | 11.176 |
| Rose Funeral | The Desolate Form | 11.244 |
| Tame Impala | Reality In Motion | 11.256 |
| S Club 7 | Two In A Million | 11.399 |
| Vanessa Carlton | Private Radio | 11.405 |

## A.6 Spectral Contrast Retrieval

**Table 6: Top 10 Retrieved Tracks using Spectral Contrast Retrieval under Top Genre Relevance**

| Artist | Song | Similarity Score |
|---|---|---|
| Dr. Hook | Sharing The Night Together | 0.999 |
| Andreas Bourani | Auf Anderen Wegen | 0.999 |
| Stereophonics | A Minute Longer | 0.999 |
| CHVRCHES | Never Say Die | 0.999 |
| The Preatures | Yanada | 0.999 |
| a-ha | Under The Makeup | 0.998 |
| Renaissance | Northern Lights | 0.998 |
| Bryan Adams | You Can't Take Me | 0.999 |
| Spice Girls | Goodbye | 0.998 |
| Thompson Twins | King For A Day | 0.998 |

## A.7 VGG19 Retrieval

**Table 7: Top 10 Retrieved Tracks using VGG19 Retrieval under Top Genre Relevance**

| Artist | Song | Similarity Score |
|---|---|---|
| Alter Bridge | Wonderful Life | 0.951 |
| Suprême NTM | Laisse pas traîner ton fils | 0.950 |
| Doves | Pounding | 0.945 |
| Jonas Brothers | Hello Beautiful | 0.944 |
| Bat for Lashes | A Wall | 0.941 |
| Theatre of Tragedy | A Rose for the Dead | 0.940 |
| Depeche Mode | Going Backwards | 0.938 |
| Lacrimas Profundere | One Hope's Evening | 0.938 |
| Josh Groban | Granted | 0.937 |
| Haken | 1985 | 0.936 |

## A.8 ResNet Retrieval

**Table 8: Top 10 Retrieved Tracks using ResNet Retrieval under Top Genre Relevance**

| Artist | Song | Similarity Score |
|---|---|---|
| Miguel Bosé | Duende | 40.754 |
| Bat for Lashes | A Wall | 41.078 |
| Mariah Carey | It’s a Wrap | 42.983 |
| Depeche Mode | Going Backwards | 43.084 |
| Casper | Hinterland | 44.339 |
| Haken | 1985 | 44.776 |
| Theatre of Tragedy | A Rose for the Dead | 44.785 |
| Alter Bridge | Wonderful Life | 45.034 |
| Snow Patrol | Somewhere a Clock Is Ticking | 45.538 |
| Lacuna Coil | Underdog | 46.074 |

## A.9 Early Fusion BERT+MFCC Retrieval

**Table 9: Top 10 Retrieved Tracks using Early Fusion BERT+MFCC Retrieval under Top Genre Relevance**

| Artist | Song | Similarity Score |
|---|---|---|
| Britney Spears | Break the Ice | 0.942 |
| Kylie Minogue | My Secret Heart | 0.939 |
| Backstreet Boys | Safest Place to Hide | 0.934 |
| Skank | Esquecimento | 0.933 |
| Heart | Dreamboat Annie | 0.932 |
| Tame Impala | Reality In Motion | 0.929 |
| Jamiroquai | Lifeline | 0.928 |
| S Club 7 | Two In A Million | 0.927 |
| Klymaxx | I Miss You | 0.927 |
| Foxy Shazam | Introducing Foxy | 0.927 |

## A.10 Late Fusion MFCC+VGG19 Retrieval

**Table 10: Top 10 Retrieved Tracks using Late Fusion MFCC+VGG19 Retrieval under Top Genre Relevance**

| Artist | Song | Similarity Score |
|---|---|---|
| Sheryl Crow | Superstar | 0.921 |
| James Morrison | One Last Chance | 0.916 |
| Lacrimas Profundere | One Hope’s Evening | 0.909 |
| Britney Spears | Break the Ice | 0.908 |
| Hey Violet | Break My Heart | 0.906 |
| Lady Antebellum | Bartender | 0.902 |
| Suprême NTM | Laisse pas traîner ton fils | 0.902 |
| Jonas Brothers | Hello Beautiful | 0.901 |
| Take That | Hold On | 0.900 |
| Laura Pausini | Bendecida pasión | 0.899 |

# B FINAL EVALUATION RESULTS OF RETRIEVAL SYSTEMS

**Table 11: Final Evaluation Results of Retrieval Systems**

| Retrieval System | Relevance Definition | Precision@10 | Recall@10 | NDCG@10 | MRR | Coverage@10 | Tag Diversity@10 | Popularity Diversity@10 | AvgPop@10 |
|---|---|---|---|---|---|---|---|---|---|
| Random Retrieval | top_genre | 0.042 | 0.002 | 0.042 | 0.098 | 99.98 | 9.72 | 190.83 | 35.11 |
| Tag-Based Retrieval | top_genre | 0.095 | 0.008 | 0.094 | 0.195 | 88.17 | 5.20 | 161.03 | 34.99 |
| TF-IDF Retrieval | top_genre | 0.083 | 0.007 | 0.096 | 0.188 | 89.87 | 9.61 | 163.02 | 34.76 |
| BERT Retrieval | top_genre | 0.088 | 0.008 | 0.089 | 0.179 | 83.27 | 9.63 | 182.73 | 37.47 |
| MFCC Retrieval | top_genre | 0.097 | 0.008 | 0.103 | 0.209 | 92.78 | 9.52 | 167.94 | 36.10 |
| Spectral Contrast Retrieval | top_genre | 0.067 | 0.006 | 0.068 | 0.150 | 92.27 | 9.63 | 179.70 | 36.05 |
| VGG19 Retrieval | top_genre | 0.077 | 0.006 | 0.085 | 0.193 | 87.55 | 9.63 | 170.17 | 36.32 |
| ResNet Retrieval | top_genre | 0.082 | 0.006 | 0.091 | 0.208 | 82.76 | 9.63 | 168.27 | 35.25 |
| Early Fusion BERT+MFCC Retrieval | top_genre | 0.098 | 0.008 | 0.103 | 0.208 | 91.06 | 9.52 | 171.06 | 36.71 |
| Late Fusion MFCC+VGG19 Retrieval | top_genre | 0.099 | 0.008 | 0.108 | 0.231 | 83.60 | 9.61 | 176.08 | 37.84 |
| Random Retrieval | tag_overlap | 0.0002 | 0.0003 | 0.0003 | 0.0005 | 100.00 | 9.73 | 191.58 | 35.10 |
| Tag-Based Retrieval | tag_overlap | 0.0010 | 0.0023 | 0.0028 | 0.0030 | 88.17 | 5.20 | 161.03 | 34.99 |
| TF-IDF Retrieval | tag_overlap | 0.0009 | 0.0023 | 0.0014 | 0.0023 | 89.87 | 9.71 | 163.02 | 34.76 |
| BERT Retrieval | tag_overlap | 0.0008 | 0.0016 | 0.0013 | 0.0025 | 83.27 | 9.63 | 182.73 | 37.47 |
| MFCC Retrieval | tag_overlap | 0.0015 | 0.0033 | 0.0027 | 0.0054 | 92.78 | 9.52 | 167.94 | 36.10 |
| Spectral Contrast Retrieval | tag_overlap | 0.0008 | 0.0014 | 0.0013 | 0.0026 | 92.27 | 9.63 | 179.70 | 36.05 |
| VGG19 Retrieval | tag_overlap | 0.0018 | 0.0047 | 0.0053 | 0.0108 | 87.55 | 9.63 | 170.17 | 36.32 |
| ResNet Retrieval | tag_overlap | 0.0018 | 0.0046 | 0.0056 | 0.0119 | 82.76 | 9.63 | 168.27 | 35.25 |
| Early Fusion BERT+MFCC Retrieval | tag_overlap | 0.0014 | 0.0030 | 0.0026 | 0.0052 | 91.06 | 9.61 | 171.06 | 36.71 |
| Late Fusion MFCC+VGG19 Retrieval | tag_overlap | 0.0025 | 0.0060 | 0.0066 | 0.0133 | 83.60 | 9.61 | 176.08 | 37.84 |