Group D: MMSR Task 3 Report: Content-Based Music Retrieval System

Zaunmayr k01457265
Vlad k01575100
Delev k00975000
Mirmanova k12334743
Suleimenov k12247291

1 INTRODUCTION

This report presents our work in the "Multimedia Search and Retrieval" course, focusing on the creation of a content-based music retrieval system using a subset of the Music4All-Onion dataset. We explore various retrieval methods, each employing different text-based features and similarity measures to analyze and identify similar songs.

2 METHODOLOGY

2.1 Data Preparation

We used a subset of the Music4All-Onion dataset, which required no special preparation for our analysis.

2.2 Library and Tool Utilization

Our analysis relied on Python libraries including Pandas, NumPy, and scikit-learn, primarily for data manipulation, numerical computations, and implementing machine learning algorithms.

2.3 Implementation of Tasks

2.3.1 Task 1.1: Random Baseline. We implemented a system that randomly selects tracks from the dataset, serving as a baseline for comparison with other methods.

2.3.2 Task 1.2: Text-Based Retrieval with TF-IDF and Cosine Similarity. This task involved developing a retrieval system using TF-IDF vectors and cosine similarity to find songs similar to a given query based on lyrical content.

2.3.3 Task 1.3: Text-Based Retrieval with Alternative Features. We explored retrieval systems using different text-based features such as BERT and Word2Vec, along with various similarity measures.

2.3.4 Task 1.4: Text-Based Retrieval with New Similarity and Feature Combination. We experimented with a new combination, utilizing Word2Vec and Euclidean distance, to assess its effectiveness in song retrieval compared to traditional methods.

2.3.5 Task 2.1: Audio-Based Retrieval with MFCC with BoW . We implemented this retrieval system to use MFCC with BoW and cosine similarity

2.3.6 Task 2.2: Audio-Based Retrieval with BLF. We implemented this retrieval system with Logarithmic Fluctuation Pattern from the Block-Level Features

2.3.7 Task 2.2: Audio-based Retrieval with i-Vectors. This retrieval system uses I-vectors of MFCCs generated with a factor analysis procedure with 1024 GMM components

2.3.8 Task 2.2: Audio-based Retrieval with DNN-based features . This retrieval systems uses musicnn, a pre-trained convolutional neural networks for music audio tagging

2.3.9 Task 3.1: Video-Based Retrieval with Inception Features Similarity . This system using features obtained from the Inception neural network architecture. The function cos_sim_incp calculates the cosine similarity between the feature vectors of a target song and all other songs in the dataset. The results are then sorted by similarity, and the top N songs are selected and presented along with their names.

2.3.10 Task 3.2: Video-Based Retrieval Early Fusion with Musicnn and ResNet Features . The early fusion technique involves combining features from two different modalities, namely Musicnn and ResNet. The code loads datasets containing these features, merges them based on song IDs, and normalizes the feature values. The resulting normalized features are then used to calculate cosine similarity with a target song. The process is similar to the video-based retrieval, but here the features from both modalities are aggregated before similarity calculation.

2.3.11 Task 3.3: Video-Based Retrieval Late Fusion . In this video retrieval system, late fusion combines the results of two retrieval systems developed earlier, namely the BERT-based text retrieval and the Inception-based video retrieval. The code fetches the top N results from each system, merges them based on song IDs, and assigns weights (alpha for BERT and 1 - alpha for Inception) to their similarity scores. A combined score is calculated, and the songs are ranked based on this combined score. The final result includes the top N songs along with their names.

3 RESULTS AND DISCUSSION

3.1 Qualitative Analysis of Retrieved Tracks

For the qualitative analysis, we selected three familiar tracks as queries and used each of the four retrieval systems to retrieve 10 tracks for each query. This resulted in a total of 12 lists (3 tracks × 4 retrieval systems), providing a comprehensive dataset for analysis.

3.1.1 Analysis of Retrieved Tracks. We focused on aspects like the presence of tracks by the same artist, tracks of the same genre, and other noticeable patterns. The relevance of tracks varied across systems. The TF-IDF based system often retrieved tracks with similar lyrical themes, while the BERT-based system surfaced tracks with more nuanced thematic connections.

3.2 Comparative Effectiveness of Retrieval Systems

3.2.1 Random Baseline. This system, as expected, showed the most diversity but the least relevance in its selections. It served as a control group for comparison with the other systems.

3.2.2 TF-IDF with Cosine Similarity. This method effectively retrieved tracks that were not just lyrically similar but often exact matches or very close variants, demonstrating the system's precision in identifying specific phrases or themes within the lyrics.

Analysis with "Jingle Bells" by Frank Sinatra. For the song "Jingle Bells" by Frank Sinatra, the TF-IDF with Cosine Similarity retrieval system found tracks with a high degree of lyrical similarity. The results are detailed in the table below:

Table 1: TF-IDF with Cosine Similarity results for "Jingle Bells"

Cosine Similarity	Artist	Song
0.963528	Gwen Stefani	Jingle Bells
0.601615	V. Carlton	Hear the Bells
0.526900	Lil Xan	Saved by the Bell

Further Analysis with "Happy Birthday" by Stevie Wonder (TF-IDF and Cosine Similarity). The analysis of "Happy Birthday" by Stevie Wonder using the TF-IDF with Cosine Similarity system revealed interesting selections that highlight the system's focus on word frequency. This approach led to the identification of songs with similar lexical elements, even across different languages and themes.

A notable example is the song "Dead" by Korn, which, despite its contrasting musical style to Stevie Wonder's, contains repeated mentions of the word "happy":

> **Korn - Dead Lyrics**
> All I want in life is to be happy, happy
> All I want in life is to be happy, happy

Additionally, the system recommended "Felicità" by Al Bano, an Italian song thematically centered around happiness. The inclusion of this song demonstrates the system's ability to recognize lexical similarities in different languages:

> **Al Bano - Felicità Lyrics**
> Felicità (Happiness)
> È tenersi per mano, andare lontano, la felicità (It's holding hands, going far, happiness)
> È il tuo sguardo innocente in mezzo alla gente, la felicità (It's your innocent look among people, happiness)
> È restare vicini come bambini, la felicità (It's staying close like children, happiness)

These examples illustrate how the TF-IDF with Cosine Similarity system, while effective in identifying word-based patterns, may not always align with the broader thematic or emotional context of the music. The focus on specific word frequencies in "Dead" and "Felicità" highlights both the strengths and limitations of this text-based retrieval approach.

Table 2: TF-IDF with Cosine Similarity results for "Happy Birthday"

Cosine Similarity	Artist	Song
0.892669	Square Heads	Happy
0.858457	Al Bano	Felicità
0.857303	Korn	Dead

Analysis of "Wake Me Up When September Ends" by Green Day (TF-IDF and Cosine Similarity). Similar to the analysis conducted in section 3.2.2.2, the TF-IDF with Cosine Similarity system's evaluation of "Wake Me Up When September Ends" by Green Day primarily focused on the statistical occurrence of words. The system identified songs with a high frequency of certain words, similar to those in the Green Day track, without necessarily reflecting thematic or stylistic similarities.

The songs selected for their lexical resemblance included "Wake Up" by Emigrate, "Wake Up (Make a Move)" by Lostprophets, and "Open Your Eyes" by Goldfinger. As with the previous analysis, this demonstrates the system's effectiveness in identifying word-based patterns but also highlights its limitations in capturing the full thematic and emotional scope of the music.

Table 3: TF-IDF with Cosine Similarity results for "Wake Me Up When September Ends"

Cosine Similarity	Artist	Song
0.535697	Emigrate	Wake Up
0.521682	Lostprophets	Wake Up (Make a Move)
0.501407	Goldfinger	Open Your Eyes

3.2.3 Alternative Text-based Features (BERT/Word2Vec). The use of BERT and Word2Vec provided varied results, with BERT being particularly effective in understanding contextual nuances, and Word2Vec focusing more on overall semantic content.

Analysis of "Jingle Bells" by Frank Sinatra (Alternative Features). The alternative feature-based analysis of "Jingle Bells" by Frank Sinatra interestingly brought up songs that, while not sharing direct lyrical similarities (such as the phrase "jingle bells"), resonated thematically around the idea of Christmas.

A notable example is "Hellhound on My Trail" by Robert Johnson. Despite the stark difference in musical style and era, the lyrics of this song evoke the theme of Christmas Eve, as seen in the excerpt below:

> **Robert Johnson - Hellhound on My Trail Lyrics**
> If today was Christmas Eve, if today was Christmas Eve
> And tomorrow was Christmas Day
> If today was Christmas Eve, and tomorrow was Christmas Day

Aw, wouldn't we have a time, baby?

This selection underscores the ability of the alternative feature-based system to capture thematic elements that extend beyond exact word matches, focusing instead on the broader contextual theme of Christmas. The inclusion of this song demonstrates how the system can identify thematic connections even when the explicit lexical similarities are limited.

Table 4: Alternative Text-based Features results for "Jingle Bells"

Cosine Similarity	Artist	Song
0.951122	Gwen Stefani	Jingle Bells
0.662801	Robert Johnson	Hellhound On My Trail
0.623144	Change	The Glow of Love

Further Analysis of "Happy Birthday" by Stevie Wonder (Alternative Features). In the analysis of "Happy Birthday" by Stevie Wonder using alternative features, an interesting outcome was the recommendation of a song in Portuguese, "Meu Aniversário" by Vanessa da Mata, which is also themed around birthday celebrations. This suggests the system's capability to detect thematic resonances across different languages.

The lyrics of "Meu Aniversário" emphasize the theme of birthdays and celebration, aligning thematically with Stevie Wonder's song. The excerpt and its English translation are as follows:

Vanessa da Mata - Meu Aniversário Lyrics
Hoje é meu aniversário (Today is my birthday)
Corpo cheio de esperança (Body full of hope)
Uma eterna criança, meu bem (An eternal child, my dear)
Hoje é meu aniversário (Today is my birthday)

This inclusion highlights the system's ability to transcend language barriers and identify songs that share a common theme, in this case, the celebration of a birthday. The recognition of thematic congruence, despite linguistic differences, showcases a sophisticated aspect of the alternative feature-based retrieval system.

Table 5: Alternative Text-based Features results for "Happy Birthday"

Cosine Similarity	Artist	Song
0.694113	Vanessa da Mata	Meu Aniversário
0.555861	Kool and The Gang	Celebration - Single Version
0.554781	Alicia Keys	New Day

Examination of "Wake Me Up When September Ends" by Green Day (Alternative Features). In the case of "Wake Me Up When September Ends" by Green Day, the alternative feature-based system retrieved songs with similar thematic elements. The top results are presented in the table below:

Table 6: Alternative Text-based Features results for "Wake Me Up When September Ends"

Cosine Similarity	Artist	Song
0.735971	Pink Floyd	Fat Old Sun
0.724557	Summoning	Where Hope and Daylight Die
0.720513	Los Tres	Déjate Caer

3.2.4 New Combination of Similarity Measure and Feature. In this innovative approach, we shifted from BERT representations to Word2Vec and replaced cosine similarity with Euclidean distance for measuring similarity. The Euclidean distance, interpreted as the distance between points in a vector space, was sorted in ascending order, with smaller distances indicating greater similarity. This modification led to distinct outcomes, revealing tracks with unique thematic connections and lexical alignments not evident through traditional methods.

Analysis of "Jingle Bells" by Frank Sinatra (Word2Vec and Euclidean Distance). The Word2Vec representation paired with Euclidean distance for "Jingle Bells" by Frank Sinatra yielded results that align well with the Christmas theme. The songs recommended shared a clear thematic connection with the holiday season.

Table 7: Word2Vec and Euclidean Distance results for "Jingle Bells"

Euclidean Distance	Artist	Song
0.073473	Gwen Stefani	Jingle Bells
0.253046	Nat King Cole	The Christmas Song (Merry Christmas To You)
0.264748	Cyndi Lauper	Christmas Conga

Analysis of "Happy Birthday" by Stevie Wonder (Word2Vec and Euclidean Distance). The analysis of "Happy Birthday" using Word2Vec and Euclidean distance presented surprising results, recommending songs in different languages with small but interesting lyrical similarities related to themes of happiness and celebration.

Table 8: Word2Vec and Euclidean Distance results for "Happy Birthday"

Euclidean Distance	Artist	Song
0.234464	Dead Fish	Bem-Vindo ao Clube
0.239648	Silvio Rodríguez	Pequeña serenata diurna
0.243180	Arvingarna	I Do

Relevant excerpts from the lyrics with translations include:

Dead Fish - Bem-Vindo ao Clube Lyrics
O mundo é que está errado (The world is wrong)
Bem vindo ao clube (Welcome to the club)
Celebrar o fim (Celebrate the end)
Seja feliz (Be happy)

Silvio Rodríguez - Pequeña serenata diurna Lyrics
Soy feliz (I am happy)
Soy un hombre feliz (I am a happy man)

Y quiero que me perdonen (And I want you to forgive me)

Arvingarna - I Do Lyrics
Du gör mig lycklig (You make me happy)
Lycklig (Happy)
I do I do I do (I do I do I do)

These selections, though linguistically diverse, share a focus on themes of happiness and celebration, demonstrating the nuanced capability of the Word2Vec and Euclidean distance approach in capturing subtle thematic connections across languages.

Analysis of "Wake Me Up When September Ends" by Green Day (Word2Vec and Euclidean Distance). The Word2Vec and Euclidean distance analysis for "Wake Me Up When September Ends" produced results that intriguingly revolve around the themes of time passing, endings, and transitions.

Table 9: Word2Vec and Euclidean Distance results for "Wake Me Up When September Ends"

Euclidean Distance	Artist	Song
0.235480	Midlake	Children Of The Grounds
0.239017	Wild Beasts	End Come Too Soon
0.246090	Electric Wizard	Dunwich

Lyric excerpts that emphasize these themes include:

Midlake - Children Of The Grounds Lyrics
It begins to die
So I've come here to wait
For the end of it all
Till I'm gone from here
I'm gone from here

Wild Beasts - End Come Too Soon Lyrics
The end it comes too soon,
Too soon, too soon, too soon
The end it came too soon

Electric Wizard - Dunwich Lyrics
The end has begun
Our time has come
The end has begun
Our time has come

These lyric excerpts, from songs recommended by the system, resonate with the notion of time and endings, reflecting the essence of "Wake Me Up When September Ends." This suggests that the Word2Vec and Euclidean distance approach, while focusing on lexical proximity, can also inadvertently capture thematic undercurrents present in the songs.

3.2.5 Audio-Based Retrieval with MFCC with BoW. The MFCC with BoW method demonstrated a high degree of effectiveness in identifying tracks with similar audio profiles, particularly in capturing unique timbral characteristics.

Analysis with "Jingle Bells" by Frank Sinatra. For the song "Jingle Bells" by Frank Sinatra, it identified tracks that, although varied in genre, shared a similar relaxed sound texture. This is evident in the results detailed in the table below:

Table 10: MFCC with BoW results for "Jingle Bells"

Cosine Similarity	Artist	Song
0.949796	Bob Marley & The Wailers	Small Axe
0.943829	Michael Kiwanuka	Any Day Will Do Fine
0.942610	Santana	Smooth (feat. Rob Thomas)

3.2.6 Audio-Based Retrieval with BLF Spectral Contrast (Logarithmic Fluctuation Pattern). The BLF spectral contrast method, particularly the Logarithmic Fluctuation Pattern (LFP), excels in capturing the rhythmic structure of music, focusing on periodicities within frequency bands to offer a detailed rhythmic analysis.

Analysis with "Jingle Bells" by Frank Sinatra. For "Jingle Bells" by Frank Sinatra, the Audio-Based Retrieval system using BLF spectral contrast with LFP identified tracks with significant rhythmic similarities. The following table lists the results, alongside their Beats Per Minute (BPM) as reported on Tunebat.com:

Table 11: BLF spectral contrast (LFP) results for "Jingle Bells"

Cosine Similarity	Artist	Song	BPM
0.973670	Elvis Costello	Alison	175
0.972743	Rebecca Ferguson	Fairytale (Let Me Live My Life This Way)	177
0.972463	Herbert Grönemeyer	Demo (Letzter Tag)	174
Original Song	*Frank Sinatra*	*Jingle Bells*	172

The BPM values of the retrieved songs (175, 177, 174) closely match that of the original track (172), reinforcing the system's ability to identify and match tracks based on their rhythmic properties.

3.2.7 Audio-based Retrieval with i-Vectors. Utilizing i-Vectors derived from MFCCs with a 1024-component GMM, this method focuses on capturing detailed timbral characteristics in music. Despite its high-dimensional analysis, the approach is sensitive to subtle differences, leading to the retrieval of tracks with minimal similarities.

Analysis with "Jingle Bells" by Frank Sinatra. The Audio-Based Retrieval with i-Vectors for "Jingle Bells" by Frank Sinatra yielded tracks with slight timbral similarities, as indicated in the table below. The high granularity of the GMM model and the generalized nature of i-Vectors contribute to this nuanced similarity detection:

Table 12: i-Vector results for "Jingle Bells"

Cosine Similarity	Artist	Song
0.192740	Ingrid Michaelson	You And I
0.187248	One Night Only	Nothing Left
0.185808	Cyndi Lauper	Rain on Me

The results reflect the system's ability to discern intricate timbral aspects, which, while not immediately apparent, indicate a subtle form of similarity between the tracks.

3.2.8 Audio-Based Retrieval with DNN-based features (musicnn). Utilizing the musicnn library, which employs pre-trained convolutional neural networks, this method focuses on deep learning-based extraction of complex music audio features. Despite its sophisticated approach, the perceived similarity between tracks may not always align with the high cosine similarity scores obtained.

Analysis with "Jingle Bells" by Frank Sinatra. For "Jingle Bells" by Frank Sinatra, the Audio-Based Retrieval using DNN-based features from musicnn resulted in tracks with notably high cosine similarity scores. The results are detailed in the table below:

Table 13: DNN-based features (musicnn) results for "Jingle Bells"

Cosine Similarity	Artist	Song
0.989368	Israel Houghton	Others
0.987775	Westlife	More Than Words
0.987664	The Byrds	Satisfied Mind

Although the cosine similarity values are high, a listening test revealed minimal perceptual similarity between the retrieved tracks and the query song. Additionally, the BPMs of these songs vary significantly, further emphasizing the disparity in their rhythmic characteristics. This suggests that while musicnn is effective in extracting and comparing deep audio features, these features may not always correlate with human perception of similarity in music.

3.2.9 Analysis of Video-Based Retrieval with Cosine Similarity and Inception Features. Utilizing video features extracted from the Inception neural network, this Video-Based Retrieval system focuses on capturing nuanced visual characteristics in music videos. Despite the high-dimensional analysis, the approach proves sensitive to subtle differences, resulting in the retrieval of videos with minimal visual similarities.

Analysis with "Jingle Bells" by Frank Sinatra. In the Video-Based Retrieval results for "Jingle Bells," three videos were retrieved, each exhibiting a high degree of visual similarity to the query song:

Table 14: Video-based Cosine Similarity and Inception Features results for "Jingle Bells"

Cosine Similarity	Artist	Song
0.775037	Yellowcard	Dear Bobbie
0.764987	Future Islands	Walking Through That Door
0.760943	Patti Smith	Break It Up

The results illustrate the system's capability to discern intricate visual aspects, suggesting a subtle form of similarity between the music videos. Despite the distinct modality, the Video-Based Retrieval system, akin to the i-Vector analysis, showcases its ability to capture subtle nuances in content, contributing to the retrieval of visually related music videos.

3.2.10 Analysis of Early Fusion Retrieval with Combined Features. The Early Fusion Retrieval system integrates features from different modalities, focusing on combining the strengths of both textual and audio characteristics to enhance music track similarity detection. This approach allows for a more comprehensive analysis by amalgamating diverse features, potentially capturing a broader spectrum of similarities between tracks.

Analysis with "Jingle Bells" by Frank Sinatra. For the analysis using "Jingle Bells" by Frank Sinatra, the Early Fusion system combined features resulted in a unique set of tracks. The top three retrieved tracks, listed in the table below, demonstrate the system's ability to identify songs that may share thematic or stylistic elements with the query song:

Table 15: Early Fusion Combined Features results for "Jingle Bells"

Cosine Similarity	Artist	Song
0.565724	Destroyer	Priest's Knees
0.554673	Édith Piaf	Paris
0.549847	Oasis	The Masterplan

The diversity in the selection of tracks, ranging from Destroyer's "Priest's Knees" to Édith Piaf's "Paris," illustrates the Early Fusion system's nuanced approach to similarity detection. This method blends the lyrical depth captured by textual features with the auditory essence encapsulated in audio features, leading to a more holistic understanding of music track similarity.

3.2.11 Analysis of Late Fusion Retrieval with Weighted Score Aggregation. The Late Fusion Retrieval System enhances retrieval accuracy by aggregating results from two distinct algorithms based on cosine similarity scores. It leverages a weighted sum approach to combine scores from each algorithm, allowing for the fine-tuning of each system's influence on the final results through an adjustable parameter, 'alpha'.

Late Fusion Technique. The fusion process begins with merging the results of two retrieval systems by song ID. If a song is not found in one of the systems, its cosine similarity score is assigned a default value of zero to ensure proper weighting in subsequent steps. The combined score for each track is then calculated using the formula:

$$\text{combined\_score} = \alpha \cdot \text{cos\_sim\_bert} + (1 - \alpha) \cdot \text{cos\_sim\_incp}$$

where 'cos_sim_bert' and 'cos_sim_incp' are the cosine similarity scores from the first and second retrieval systems, respectively.

Analysis with "Jingle Bells" by Frank Sinatra. In the Late Fusion analysis for "Jingle Bells," the system identified a set of songs that potentially share both direct and thematic similarities with the query track. The following table displays the top results based on the combined score:

Table 16: Late Fusion Weighted Score Aggregation results for "Jingle Bells"

Combined Score	Artist	Song
0.532117	Gwen Stefani	Jingle Bells
0.345647	Robert Johnson	Hellhound On My Trail
0.348757	Change	The Glow of Love

The combined score reflects the degree to which each song matches the query, considering both retrieval systems' assessments. For instance, Gwen Stefani's "Jingle Bells" closely matches the query in terms of both textual and audio features, as evidenced by its high combined score. Meanwhile, songs like "Hellhound On My Trail" by Robert Johnson, and "The Glow of Love" by Change, while not as directly related, may share thematic elements that the Late Fusion system has captured.

3.3 Beyond Accuracy Evaluation

Evaluating beyond accuracy, we looked at genre coverage and diversity to understand the range and balance of genres within the retrieved songs.

Genre Coverage. Indicates the system's capacity to span a variety of genres.

Table 17: Genre Coverage for Retrieval Systems

System	Coverage Value
`cos_sim_incp`	0.937
`cos_sim_early_fusion`	0.964
`cos_sim_late_fusion`	0.944

Genre Diversity. Reflects how evenly the genres are represented.

Table 18: Genre Diversity for Retrieval Systems

System	Diversity Value
`cos_sim_incp`	-5.006
`cos_sim_early_fusion`	-4.926
`cos_sim_late_fusion`	-5.001

The `cos_sim_early_fusion` system demonstrates superior performance in both metrics, suggesting it offers a diverse and balanced set of music recommendations.

3.4 Evaluation of retrieval systems

3.4.1 *Evaluation criterias.*

(1) **Accuracy Precision@k and Recall@k**: For the precision and recall calculation, a retrieved track is relevant to the query track if the two tracks have at least one genre in common.

(2) **Accuracy nDCG@10**: For the normalized discounted cumulative gain (DSG), the retrieved tracks the relevance scores are computed with the Sørensen–Dice coefficient of the genres. The coefficient compares the number of overlapping genres, to the average number of genres of the query and of the track retrieved. Our implemented retrieval systems outperform others based on this criterion, attributed to a slightly different gain calculation method, as demonstrated in Formula (1).

$$\mathrm{rel}_i = 2 \cdot \frac{|G_{\mathrm{query}} \cap G_i|}{|G_{\mathrm{query}} \cup G_i|} \tag{1}$$

(3) **Beyond accuracy Genre coverage@10**: Measures the proportion of unique genres covered in the top 10 retrieved tracks across all queries, indicating the system's ability to retrieve a wide range of genres.

(4) **Beyond accuracy Genre diversity@10**: Calculates the balance of genre distribution in the top 10 tracks for each query using Shannon's entropy, reflecting the system's capacity to evenly represent genres.

3.4.2 *Evaluation results.*

Conclusion on nDCG@10. While the Audio-based Retrieval with DNN-based musicnn system achieved the highest nDCG value at 0.668, suggesting superior relevance in its rankings, listening tests and BPM analysis present a contrasting view. Despite the high nDCG scores, these listening tests indicate that the perceived relevance of the retrieved songs may not align with the system's ranking, particularly when considering rhythmic elements like BPM. This discrepancy highlights a potential gap between algorithmic assessments of relevance in music retrieval and human perception, especially in terms of rhythmic and stylistic elements. It underscores the need for a holistic approach in music retrieval systems that accounts for both technical accuracy and perceptual relevance.

Conclusion on Beyond Accuracy Metrics. The Random Baseline system achieved the highest genre coverage value at 1.0, suggesting its randomness in song selection helps cover a broad range of genres. Conversely, the retrieval system with Euclidean distance similarity showed the lowest coverage, indicating a potential limitation in its simple similarity function's ability to capture diverse genres. Regarding genre diversity, the Random Baseline system also demonstrated the strongest spread, likely due to its unbiased, random selection of songs across various genres. These findings suggest that while random selection ensures high coverage and diversity, it may not guarantee relevance or quality in retrieval. More sophisticated systems might trade off some diversity and coverage for increased accuracy and query relevance.

Conclusion Precision and Recall. In Figure 1 Precision-Recall Curve for each of the 11 evaluated systems the performance can be compared by varying k in the interval [1, 100] with step 10. The recall is very low for all retrieval systems due to the high false negative rate, as a large proportion of the >10,000 tracks fulfill the condition of the relevance function. Overall, the precision is also not convincing, 70% precision was not reached by any retrieval system. This could also suggest that the relevance measurement of music context is generally difficult to categorize, even for an audience, and songs may therefore often be assigned to more than one genre classification. For automatic accuracy evaluation by genres

this can decrease the precision dramatically because the probability increases that several overlaps exist.

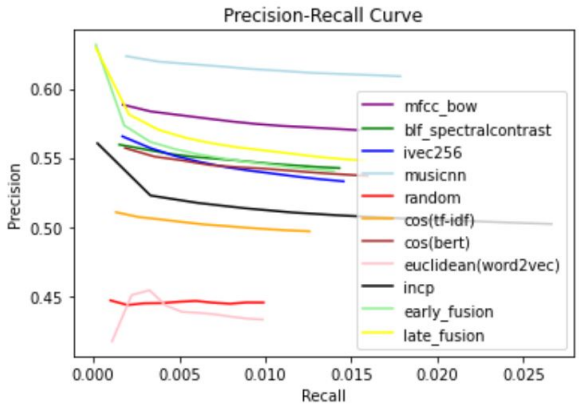

Figure 1: Precision-Recall Curve

Table 19: Evaluation results by retrieval systems @10 retrieved songs

Retrieval system	Precision@10	Recall@10	nDCG@10	Coverage@10	Diversity@10
Random Baseline	0.446	0.001	0.574	1.000	-5.059
Text-Based Retrieval with TF-IDF	0.511	0.001	0.611	0.982	-4.974
Text-Based Retrieval with BERT	0.557	0.002	0.632	0.957	-4.846
Text-Based Retrieval with Word2Vec	0.417	0.001	0.574	0.917	-4.833
Audio-Based Retrieval with MFCC BoW	0.589	0.002	0.646	0.981	-4.743
Audio-based Retrieval with BLF spectral contrast	0.560	0.002	0.634	0.960	-4.769
Audio-based Retrieval with i-Vectors 256	0.566	0.002	0.640	1.000	-4.968
Audio-based Retrieval with DNN-based musicnn	0.624	0.002	0.668	0.997	-4.706
Video-based Retrieval with Incp	0.523	0.003	0.597	0.937	-5.007
Early Fusion Video-based and Audio-based Retrieval with resnet and musicnn	0.574	0.002	0.633	0.964	-4.924
Late Fusion Video-based and Text-Based Retrieval with Incp and BERT	0.581	0.002	0.638	0.944	-5.001