Transformer-Based Multimodal Framework for Music Similarity Analysis and Recommendation Systems

Mikhail Rumiantcev
University of Jyväskylä
Jyväskylä, Finland
mikhail.rumiantcev@jyu.fi

***Abstract*—Modern music streaming platforms offer vast catalogs and personalized discovery experiences. Nevertheless, current music recommendation systems often overemphasize popular content and fail to capture complex user preferences or support the exploration of niche genres. This paper addresses these limitations by proposing a deep learning–based multimodal recommendation framework that leverages transformer architectures to analyze audio signals and contextual metadata. The goal is to enhance music similarity modeling and recommendation accuracy by generating enriched embeddings that capture musical structure, instrumentation, and genre subtleties. The research introduces a system that combines audio-based features with metadata through a fusion strategy informed by attention mechanisms. The methodology includes large-scale experimentation on public music datasets and evaluation using standard recommendation quality metrics. Results demonstrate improved personalization and diversity in recommendations compared to baseline models. This work contributes to the field by providing a novel multimodal architecture and demonstrating the effectiveness of audio-content-aware recommendation strategies.**

I. Introduction

Music streaming services today provide access to vast and diverse catalogs, allowing users to explore various artists, genres, and styles. Music recommendation systems commonly employ collaborative filtering, content-based filtering, or hybrid approaches that combine both to assist in the discovery process. However, despite their success in improving personalization, several open challenges remain. For instance, Velankar and Kulkarni (2022) [1] highlight issues such as cold start problems, insufficient data, and unreliable suggestions that limit the effectiveness of current systems. These limitations hinder the ability of recommendation systems to deliver personalized, relevant content, impacting user engagement and satisfaction.

Personalization, diversity, novelty, and contextual relevance remain critical challenges in music recommendation systems, as existing models often struggle to balance these factors effectively. Furthermore, integrating multimodal data—such as audio features, user interaction history, social influence, and textual metadata—introduces significant computational and methodological complexities. Despite ongoing research, it is still unclear how future music recommendation systems can utilize multimodal machine-learning techniques to enhance personalization and user satisfaction. Overcoming these challenges is crucial for building more effective and engaging music recommender systems. Perera et al. (2020) [2]

Music similarity research is fundamental for recommendation systems, as it identifies tracks, artists, or genres with comparable attributes, thereby facilitating the suggestion of music that aligns with a user's preferences. The application of transformer-based architectures in music recommendation remains underexplored, particularly in multimodal fusion and similarity modeling. Existing systems rarely exploit the capacity of self-attention mechanisms to model long-range dependencies in audio or to align semantic information across modalities. As a result, recommendations often reflect biases toward popular content and fail to surface niche or culturally diverse tracks. Additionally, many current approaches lack interpretability, making tracing or justifying similarity-based decisions difficult. These challenges motivate the development of a unified, transformer-based multimodal framework that integrates audio and metadata to enhance music similarity analysis. Such a framework should not only improve recommendation accuracy and diversity but also enable structured, ontology-based reasoning to increase the transparency and controllability of the recommendation process.

The primary goal of this research is to develop a unified transformer-based multimodal framework that enhances music similarity modeling and recommendation quality by integrating audio content and contextual metadata. The study aims to design a transformer architecture capable of capturing long-range dependencies in audio signals using self-attention mechanisms, enabling a more accurate representation of musical structure and characteristics. A key objective is to develop an effective multimodal fusion strategy that aligns audio features with contextual metadata, resulting in enriched joint embeddings. The study also seeks to address the issue of popularity bias by improving similarity analysis to support the discovery of niche and culturally diverse music. Furthermore, the research introduces an ontology-driven similarity layer based on learned embeddings to provide a semantically structured and interpretable basis for recommendations. The proposed approach will be evaluated on public music datasets using established metrics to assess accuracy, diversity, and novelty improvements compared to existing baseline models.

The structure of this paper is designed to effectively guide the reader through the research. The next section delves into related work in the field, establishing a clear context for the research problem. The third section outlines the methodology and models utilized in this study, followed by a thorough description of the experimental setup and evaluation in Section 4. Finally, Section 5 presents the results of the experiments,

critically discuss the limitations of the approach, and proposes promising avenues for future research.

II. Related work

The proliferation of music streaming platforms has intensified the demand for robust and contextually aware music recommendation systems. Traditional approaches including collaborative filtering, content-based filtering, and hybrid techniques have formed the foundation of personalized recommendation engines. However, these methods continue to exhibit critical limitations, including the cold-start problem, popularity bias, low diversity in recommendations, and inadequate interpretability. Deldjoo et al. (2024) [3]

Collaborative filtering in recommendation systems tends to favor popular artists, leading to popularity bias and reducing the visibility of lesser-known musicians [4], [5], [6]. Despite various efforts to address this issue, it persists, as systems often prioritize well-known artists over emerging ones. The cold start problem, noted by Vr and Pillai (2018) [7], complicates recommendations for new users or newly added music due to limited interaction data. Additionally, many algorithms exhibit cultural bias by favoring recognized genres and neglecting diverse musical traditions. Hesmondhalgh et al. (2023) [8] pointed out a lack of transparency and fairness in the design of these systems on music streaming platforms. Identifying music that aligns with personal preferences can be challenging, and users may struggle to find songs or artists that match their mood. Hosey et al. (2019) [9] explored user search behavior on music platforms, emphasizing the need for improved search experiences tailored to the domain to enhance user satisfaction and efficiency.

Content-based filtering utilizes music information retrieval techniques to analyze low-level audio features, but it struggles with semantic understanding and the connection between raw signal processing and human perception [10]. Traditional systems focus on low-level features like pitch and rhythm, which may not fully capture the nuances of musical similarity [11]. Modern techniques now include mid-level and high-level features, such as harmonic progression and emotional tone. For instance, convolutional neural networks analyze spectrograms, providing deeper insights into a song's structure and texture [12].

Hybrid approaches integrating audio-based features and metadata, such as genre, artist, and year, offer a more comprehensive understanding of music similarity by incorporating multiple data dimensions. Audio features, like tempo, rhythm, and timbre, provide detailed insights into the acoustic properties of songs. At the same time, metadata captures contextual information such as the song's historical background, cultural significance, and categorization within specific genres. Bevec et al. (2024) [13] in their study combined these elements into hybrid systems for more effective capturing of the nuances of music taste, improving the accuracy and personalization of music recommendations. This approach also allows for a richer exploration of music that goes beyond the limits of individual data types, making the system adaptable to various user preferences and diverse music databases. Quadrana et al., 2018 [14] introduced matrix factorization techniques enhanced by neural networks that have improved scalability and personalization in recommendation systems. Zhang et al. (2017) [15] leveraged hybrid models for user interaction data and audio features to reduce popularity bias and enhance diversity in recommendations.

Emotion-based music recommendation represents a growing area of research, aiming to align song recommendations with user moods and contexts. Advances in affective computing have enabled systems to classify emotions using techniques such as sentiment analysis of lyrics, facial recognition, and brain-computer interfaces. Gu et al. (2020) [16] explored advancements in brain-computer interfaces, focusing on signal-sensing technologies and computational intelligence techniques such as fuzzy models and transfer learning. These methods enable effective monitoring of cognitive states during tasks, benefiting healthcare applications and research. Aruna et al. (2021) [17] utilized facial recognition technology to adapt recommendations based on real-time emotional cues dynamically. These methods demonstrate the potential of integrating physiological and contextual data into recommendation systems, offering more intuitive and adaptive user experiences. However, challenges remain in standardizing emotion classification and balancing computational efficiency with real-time responsiveness.

TABLE I. Summary of limitations in existing music recommendation systems and contributions of this study

Challenge in existing work	Identified limitations	Planned contribution of this study
Cold-start problem and lack of diversity	Existing methods struggle with new users or tracks and often generate homogenous recommendations [3], [7].	Leverage rich audio-content and semantic metadata fusion to support cold-start scenarios and enhance diversity.
Lack of Explainability and Transparency	It is difficult to justify recommendations or trace decision logic [8].	Incorporate an ontology-based reasoning layer to support interpretable recommendations.
Popularity and cultural bias	Recommender systems disproportionately favor mainstream or Western music, limiting exposure to diverse or niche content [4], [8].	Introduce balanced multimodal embeddings and diversity-aware modeling to promote underrepresented music.
Low level feature dependence	Traditional content-based filtering relies on shallow audio descriptors, which poorly reflect musical semantics [10], [11].	Employ transformer models to extract high-level, temporally aware representations from audio.
Fragmented or shallow multimodal fusion	Current multimodal systems inadequately align audio, metadata, and contextual signals [13], [14].	Proposes attention-based multimodal fusion to integrate and align heterogeneous data with semantic precision.

While prior studies have advanced the development of music recommendation systems through collaborative filtering, content-based approaches, and hybrid models, persistent challenges remain in personalization, diversity, cultural representation, and system transparency. These limitations are further compounded by difficulties in modeling long-term user behavior, integrating multimodal information, and reducing popularity bias. This study proposes a transformer-based multimodal framework that leverages deep contextual learning and structured semantic reasoning to address these gaps. Table 1 summarizes the primary shortcomings of existing approaches alongside the specific research objectives of the present work, thereby clarifying its methodological and scientific contributions to the field.

III. Methodology

This chapter outlines the methodology for developing a transformer-based multimodal framework for music similarity analysis and ontology-driven recommendation. It covers the system architecture, multimodal data representation, model training procedures, and the construction of a structured music similarity ontology that enables semantically enriched recommendation logic. Emphasis is placed on the fusion of audio and metadata embeddings and the role of ontological reasoning in improving recommendation transparency and interpretability. Figure 1 presents a high-level overview of the proposed architecture, highlighting key components and their interactions.

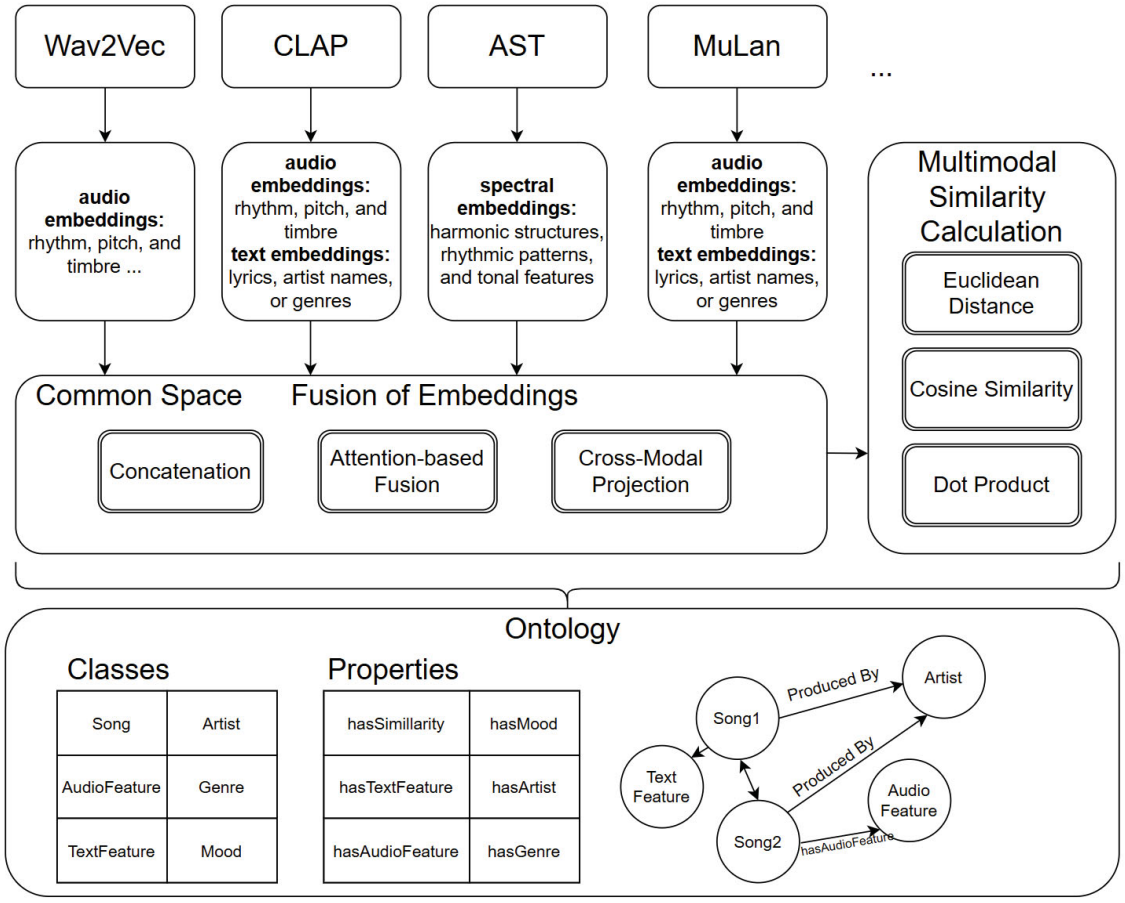

Fig.1 Transformer-based multimodal architecture

A. Transformers

Transformer-based model was originally introduced by Vaswani et al. (2017) [18] and has been used for language processing tasks. Huang et al. (2019) [19] in their research, incorporated transformer models for music processing. They discuss how transformers, with their self-attention mechanism, can be employed for tasks such as music similarity, classification, and retrieval. By leveraging the self-attention mechanism, the model effectively learns long-term dependencies in music sequences, allowing for accurate representation of melodic, harmonic, and rhythmic structures.

By leveraging the self-attention mechanism, the model effectively learns long-term dependencies in music sequences, allowing for accurate representation of melodic, harmonic, and rhythmic structures. Detailed mathematical derivations of the transformer architecture, including the attention mechanism, multi-head attention, and feedforward networks, can be found in the work by Thickstun et al. (2021) [20]. This research provides an in-depth explanation of the transformer model's mathematical

operations, such as calculating attention weights, aggregating values, and applying residual connections and layer normalization. These components form the core of the transformer's ability to model complex relationships in data, which is particularly useful in tasks like music similarity and recommendation.

The primary reason for using transformers in studying music similarity is their versatility and capability to learn complex representations from data. The model learns patterns based on fixed features in traditional approaches, such as classical machine learning methods or convolutional neural networks [21]. For example, one might use features like tempo, key, or pitch, which are manually engineered. While this is effective, it cannot adaptively discover new, more abstract patterns. Transformers automatically learn hierarchical and contextual relationships from raw data through the attention mechanism. This allows them to process music in a way that accounts for local and global structures. For instance, a transformer model could learn how a melody progresses over time, capture harmonic changes, and even recognize recurring patterns that are often difficult to define with traditional feature extraction methods. As a result, transformers can produce more accurate and generalizable representations of music that are not limited by preconceived notions of what constitutes similarity.

1) Wav2Vec: is an advanced deep learning model [22] developed by Meta AI [23] for self-supervised learning on speech data, utilizing raw audio waveforms to learn effective speech representations without extensive labeled data. The model's architecture includes a convolutional neural network for feature encoding, a transformer-based context network, and a quantization module that discretizes the latent speech representations. During pretraining, Wav2Vec 2.0 masks random sections of the input audio and trains the model to predict the masked regions using the surrounding context. This innovative approach allows the model to capture local and global dependencies in the audio signals, producing robust representations that generalize well across various speech tasks. Fine-tuning the model on labeled datasets tailors it for specific applications such as automatic speech recognition, speaker identification, and speech emotion recognition. Wav2Vec 2.0 has achieved state-of-the-art performance in ASR, significantly reducing the need for labeled data. However, its substantial computational requirements for pretraining and fine-tuning present challenges for accessibility. Nevertheless, Wav2Vec 2.0 significantly advances speech processing, opening new avenues for self-supervised learning in audio and other domains.

2) Audio Spectrum Transformer (AST): is an advanced neural network architecture [24] designed specifically for audio classification and related tasks. It operates directly on spectrogram representations of audio signals, treating them as image-like inputs. This approach enables the model to capture local and global dependencies in the audio data, making it highly effective for tasks requiring a nuanced understanding of spectral patterns. The core of the model consists of a multi-head self-attention mechanism, which allows the model to dynamically focus on different parts of the input spectrum, facilitating the recognition of complex acoustic events. Pretraining strategies, such as masked spectrogram patch prediction, are often employed to enable self-supervised learning, reducing the reliance on labeled data. Fine-tuning on downstream tasks, such as environmental sound classification or music genre identification, has shown the AST to achieve state-of-the-art performance across multiple audio benchmarks.

3) MuLan: is a multimodal transformer-based model [25] designed for understanding and generating music by aligning audio and natural language. By leveraging a contrastive learning framework, MuLan learns joint embeddings of music and text, enabling tasks such as music retrieval via text descriptions and text-based music classification. Trained on large-scale music and associated metadata datasets, it uses a dual-encoder architecture where one encoder processes audio features such as spectrograms and the other processes textual input. MuLan's ability to bridge audio and text modalities is a powerful tool for cross-modal music understanding and recommendation systems.

4) Contrastive Language-Audio Pretraining (CLAP): is an innovative framework [26] that unifies audio and textual modalities through self-supervised learning, enabling robust multimodal representation learning for audio tasks. Inspired by the success of vision-language models like CLIP [27], CLAP employs dual-encoder architecture where one encoder processes audio inputs and another processes textual descriptions. These encoders are trained jointly using a contrastive loss function to maximize the similarity between corresponding audio-text pairs while minimizing the similarity of mismatched pairs. This approach allows CLAP to learn semantically meaningful and generalizable representations that align audio signals with their natural language descriptions. The framework has been shown to excel in various downstream tasks, including zero-shot audio classification, audio captioning, and audio-based retrieval, where textual prompts can be used to query audio databases. CLAP's reliance on large-scale, diverse datasets during pretraining ensures that its representations are broadly applicable across domains, from environmental sound recognition to music analysis. However, its performance is influenced by the quality and diversity of the pretraining data, and its computational demands may limit accessibility. Despite these challenges, CLAP represents a significant step forward in bridging the gap between auditory and linguistic modalities, paving the way for more intuitive and flexible audio understanding systems.

5) RoBERTa: is a transformer-based model that employs dynamic masking during pretraining, uses larger batch sizes, and trains on significantly more data. RoBERTa demonstrates robust generalization capabilities in diverse text-based tasks like classification, summarization, and translation. Its optimizations make it a state-of-the-art model for text representation learning, widely used in research and real-world applications. MusicBERT [30] learns contextual relationships between musical events such as notes, durations, and velocities. Its architecture is adapted to capture music's hierarchical and sequential nature, making it practical for melody continuation, chord prediction, music classification, and similarity analysis. MusicBERT's specialized focus on symbolic music data

enables it to outperform general-purpose models in music-specific tasks, providing a powerful tool for music information retrieval and generation.

B. Input Modalities and Feature Extraction

The proposed model utilizes multiple input modalities to effectively capture diverse dimensions of music similarity, each contributing distinct information that aids in accurate music recommendation and analysis. These modalities include audio features, lyrics features, metadata, and user interaction data, each processed through specialized techniques to extract meaningful representations. Each modality is processed independently to extract the relevant features before being integrated into a unified model. The feature extraction process ensures that each modality contributes its unique perspective on the music, capturing a holistic view of similarity.

1) Audio Features: Audio signals provide essential information for understanding musical content. Features such as spectrograms [31], Mel-Frequency Cepstral Coefficients (MFCCs) [32], and raw waveforms are utilized to characterize the audio content. These features are typically extracted using state-of-the-art models such as VGGish[33], OpenL3 [34], or Wav2Vec2, which are designed to transform raw audio into compact, informative representations. Additionally, pre-trained models like CLAP can generate high-level audio embeddings that capture both low- and high-level musical structures, enriching the model's ability to discern subtle similarities between tracks.

2) Lyrics Features: The lyrics of a song provide insight into its thematic and emotional content, which can significantly influence listener preferences and music similarity. To transform lyrics into useful features, transformer-based natural language processing [35] models such as BERT [36], GPT [37], or SBERT [38] are utilized to create dense embeddings that capture semantic and syntactic information. Beyond simple embeddings, sentiment analysis and topic modeling can be applied to further elucidate the lyrics' emotional tone and central themes. These features provide an additional layer of understanding regarding how songs might be similar in lyrical content, contributing to more comprehensive recommendations.

3) Metadata Features: In addition to audio and lyrics, metadata - such as genre, mood, instrumentation, and artist-specific attributes - plays a crucial role in understanding music similarity. These features can be extracted from widely used datasets such as the Spotify API [39] or the Million Song Dataset [40], providing rich and structured music information. The inclusion of metadata enables the model to consider contextual information that may not be directly captured in the audio or lyrics but is still significant for accurate similarity comparisons. For example, songs within the same genre or from the same artist may be more likely to share similarities, even if their audio features differ.

C. Transformer-Based Multimodal Fusion

In music recommendation, this mechanism supports cross-modal alignment - such as between lyrical themes and audio textures - by enabling each token such as a beat, a lyric word, or a genre tag to attend to all others in the sequence. Multi-head attention further enriches this capability by allowing the model to learn diverse relational subspaces, particularly important when fusing disparate inputs like timbral features and lyrical semantics.

Two modalities: audio and text are represented as $X_a$and X_t respectively. The attention calculation mechanism is defined as:

$$\text{Attention}(Q,K,V) = \left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (1)$$

where Q, K and V are the query, key and value metrics derived from the input embeddings, and d_k is the dimensionality of key vectors.

$$MultiHead_a(Q_{a,}, K_a, V_a) = \text{Concat}(A_1, \dots, A_h)W_a^O$$
$$MultiHead_t(Q_t, K_t, V_t) = \text{Concat}(A_1, \dots, A_h)W_t^O \quad (2)$$

where h is the number of attention heads and W^O is the learned output weight matrix that projects the concatenated results back to the desired output dimension for audio and text modalities.

To enable context-aware fusion, dynamic weights are assigned to the multi-head attention outputs from each modality, based on their relevance to the current input:

$$Fusion(X_a, X_t) = \alpha_a \cdot MH_a + \alpha_t \cdot MH_t \quad (3)$$

where $\alpha_a$and α_t are the learned weights that dynamically adjust based on the input, indicating the importance of each modality in the context of the song.

The dynamic weights $\alpha_a$and α_t can be computed based on the relevance of each modality in each context, which could be learned via a softmax function over context features, such as song genre or temporal characteristics:

$$\alpha_a, \alpha_t = softmax\ (W_{context} \cdot f(X_{context}) \quad (4)$$

where $f(X_{context})$ is a function that encodes contextual features such as genre, song structure. $W_{context}$ is the weight matrix learned for context-based weighting.

The final fused representation, which will be used for similarity computation or recommendation tasks, is obtained from the fusion process:

$$Representation_{final} = Fusion(X_{audio}, X_{text}) \quad (5)$$

This representation will be dynamically adjusted based on the context of the input song, allowing the model to prioritize the most relevant features for accurate recommendation.

Various fusion techniques can be employed to integrate information from multiple modalities effectively. These methods combine the embeddings generated by different models, such as Wav2Vec2, CLAP, AST, and MuLan, into a unified representation that captures the complex relationships between audio, lyrics, and metadata. The fusion process not only preserves the distinct characteristics of each modality but also facilitates the extraction of more comprehensive and meaningful features.

Several strategies for multimodal fusion are outlined below, each with its strengths in enhancing the model's ability to capture the nuances of music similarity and recommendation.

1) Concatenation-Based Fusion: is a fusion method [41] where feature vectors from different models are combined into a single high-dimensional representation. Shi et al. (2024) [42] explored various approaches to multimodal fusion in their research. This method preserves distinct characteristics from both audio and text, enhancing the model's ability to capture multimodal relationships. For example, if an audio model Wav2Vec2 generates a 512-dimensional embedding and a joint audio-text model CLAP produces a 256-dimensional embedding, concatenating them results in a 768-dimensional vector that retains information from both domains.

2) Attention Feature Fusion: is a more adaptive fusion technique [43] that involves multi-head attention mechanisms, which assign varying importance to different modality embeddings based on context. This approach lets the system dynamically focus on the most relevant features, optimizing similarity computation based on the song's characteristics. The model may prioritize audio features in instrumental music, capturing spectral and temporal properties. For lyrically rich songs, attention may shift towards textual embeddings, emphasizing semantic and linguistic elements. This context-aware fusion enhances the model's ability to generate robust and meaningful representations across diverse music styles.

3) Cross-Modal Projection: To achieve a unified latent space, embeddings from different sources can be projected to a standard representational format using a linear transformation or a neural network-based mapping function. Aligning embeddings in the same dimensional space allows for more effective cross-modal interaction, improving the system's ability to compare and relate information from audio, lyrics, and metadata coherently.

4) Multimodal Similarity Calculation: After combining the embeddings, the system calculates the similarity between music entities to make recommendations. Cosine Similarity [44] quantifies the cosine of the angle between two vectors and is typically used when embeddings are normalized to unit length. Euclidean Distance [45] calculates the straight-line distance between two vectors in the embedding space. Dot Product [46] can assess proximity when embeddings are transformed into a space where larger values indicate more significant similarity.

D. Music similarity ontology

To define and organize the concepts, relationships, and attributes related to the similarity between music elements such as songs, genres, artists, and audio features. It is structured using a set of predefined classes and properties in an ontology. Web Ontology Language (OWL) provides a semantic framework for understanding how music items relate to one another in terms of similarity. Korzun et al. (2018) [47] demonstrated that ontological modeling and semantic interoperability enhance cultural information retrieval and user experience in ontology-enabled recommender systems. This aligns with the current study's use of ontologies for organizing multimodal music metadata, highlighting the importance of semantic technologies in developing adaptive recommender systems across cultural domains.

To integrate transformer-based models into an OWL Ontology, we need to define the classes and properties that represent key concepts in the music domain. Classes: Song; AudioFeature; TextFeature; Artist; Genre; Mood. Properties: hasGenre; hasArtist; hasAudioFeature; hasTextFeature; hasTextFeature; hasMood.

Each model's output embeddings should be mapped into the OWL ontology. This involves creating instances of classes and assigning properties that represent the features learned by each transformer. An instance of the Song class should be created for each song in the model process. Create corresponding instances in the AudioFeature and TextFeature classes for each audio and text feature. AudioFeature1: An instance of the AudioFeature class corresponding to the embedding from Wav2Vec2. TextFeature1: An instance of the TextFeature class corresponding to the text embedding from CLAP.

After extracting the features and mapping them to the appropriate classes, OWL object properties define the relationships between the entities within the ontology. For example, the Song instance is linked to its corresponding AudioFeature and TextFeature instances, such as Song1 having the AudioFeature AudioFeature1 and the TextFeature TextFeature1. In the case of multimodal models like MuLan, both audio and text relationships are established together, where Song1 is connected to AudioFeature1 and TextFeature1. Additionally, the properties hasGenre and hasArtist establish connections between the song and its genre and artist, respectively, linking Song1 to Genre1 and Artist1.

The proposed approach makes several key assumptions about the data and the nature of the problem. First, it assumes that high-quality embeddings can be generated from each modality, which is sufficiently informative to distinguish between different music tracks. Second, it assumes that the embeddings are aligned in a common space where multimodal fusion can occur effectively. The model also assumes that the various features (audio, text, and metadata) can be meaningfully related to each other in the context of similarity with music. Additionally, the methodology assumes that cross-modal relationships, such as the relationship between lyrics and melody, can be captured through the proposed attention mechanism and projection techniques.

The proposed design fully addresses the requirements through its multimodal fusion approach. By leveraging transformer-based architectures, the system effectively learns inter- and intra-modal relationships. The concatenation-based fusion ensures that no critical features are lost during integration. At the same time, multi-head attention allows the model to focus on the most relevant features in different contexts, whether for instrumental or lyrical tracks. The cross-modal projection guarantees that the embeddings from different sources can be aligned into a unified latent space, facilitating coherent comparison and similarity calculation. Finally, using multiple similarity computation methods (Cosine Similarity, Euclidean Distance, and Dot Product) ensures the system can

calculate similarity effectively across various music embeddings, providing accurate and meaningful recommendations. The system can provide robust, context-sensitive, and scalable music recommendations by solving these critical requirements.

IV. Experimental Setup and Results

This section describes the experimental setup used to evaluate the performance of the transformer-based multimodal architecture for music similarity tasks and presents the results of these experiments. The main goal is to assess the system's effectiveness in identifying music similarity across audio and textual features and evaluate how well the music similarity ontology enhances music recommendations.

A. Dataset

A dataset must adequately represent the diversity of musical genres, styles, and cultures to ensure the model generalizes well across various use cases. Transformers are data-hungry models that benefit from large-scale datasets. However, the dataset's size should be balanced against the availability of computational resources. Longer tracks allow for better modeling of temporal dependencies but require careful handling to manage memory constraints. The selection and preprocessing of an appropriate number of tracks should be a crucial component of the data preparation. Metadata such as genre labels, tempo, mood, or structure enhances supervised tasks. The quality of the audio files in the dataset is crucial, as noisy or low-resolution data can degrade model performance. The quality of audio files in a dataset is critical, as loud or low-resolution recordings can significantly impair model performance. However, high sampling rates are generally unnecessary for speech-based tracks, as speech signals can be effectively captured at lower resolutions without compromising intelligibility or feature extraction. Ethical considerations are paramount in dataset selection. Using datasets with clear licensing terms that permit academic use is essential. Datasets with permissive open-source licenses are preferred.

1) Million Song Dataset (MSD): is a large-scale benchmark dataset [40] designed to advance music information retrieval and audio analysis research. The dataset contains metadata and precomputed audio features for one million contemporary songs spanning various genres, artists, and years. The MSD does not include raw audio files but provides detailed information such as song identifiers, artist names, release years, and features like timbre, tempo, loudness, and pitch. This dataset enables scalable research by combining audio content analysis with associated metadata and user-tagged information. The MSD has been extensively used for genre classification, artist similarity, recommendation systems, and temporal analysis of musical trends. However, its reliance on precomputed features and the absence of raw audio data presents limitations for researchers aiming to develop new feature extraction methods.

2) GTZAN: is te is a widely used dataset [48] for music genre classification and audio analysis tasks. It comprises 1000 audio tracks, each 30 seconds long, spanning 10 distinct genres: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. The dataset is balanced, with 100 tracks per genre, making it an attractive choice for machine learning and signal processing research. The audio files are sampled at 22,050 Hz in monaural format, facilitating ease of computational analysis. Despite its popularity, GTZAN has limitations, including duplicate tracks, recording artifacts, and potential mislabeling of genres, which have led to calls for caution in its use and evaluation of results. Nevertheless, it remains a foundational resource in music informatics, inspiring advancements in feature extraction, classification algorithms, and deep learning applications for audio processing.

3) FSD50K: is a large-scale, open-access dataset [49] designed explicitly for sound event detection and audio tagging tasks. Published under a Creative Commons license [50], this dataset provides a rich resource for research and development in machine listening, enabling advancements in automatic sound recognition. It is characterized by its breadth of sound categories, high-quality annotations, and robust design to support diverse machine-learning tasks.

4) Free Music Archive: is a comprehensive resource [] for research in music information retrieval and related fields. It consists of over 100,000 tracks spanning various musical genres, including rock, electronic, hip-hop, classical, and more, all freely available under Creative Commons licenses. The dataset is organized into multiple subsets of varying sizes: small, medium, large, and complete, according to different computational needs and research objectives. Each track is enriched with extensive metadata, such as artist, album, genre, and release year, providing valuable context for tasks such as genre classification, music recommendation, and music similarity analysis. Additionally, FMA offers precomputed audio features, including spectral, tonal, and rhythmic descriptors, enabling researchers to focus on high-level tasks without the overhead of feature extraction. A unique strength of FMA lies in its openness and standardization, which have made it a benchmark dataset for evaluating machine learning models in MIR. However, challenges remain, such as class imbalance across genres and the potential for cultural bias due to the predominance of Western music. Despite these limitations, the FMA dataset continues to be a cornerstone for advancing the state of the art in computational music analysis. It has inspired many applications, from automatic playlist generation to audio-based emotion recognition.

B. Data Preparation

Selecting the ideal sample duration for music similarity tasks using transformer models necessitates a compromise between computational efficiency and contextual depth. Short samples (1-5 seconds) are excellent for intricate analysis, such as rhythm, but they do not provide enough context for overarching structures. Medium samples (10-30 seconds) effectively balance transitions and patterns, maintaining computational efficiency. Long samples (30+ seconds) deliver a thorough representation but pose significant computational difficulties. Generally, transformer models perform best with 10-30-second samples, effectively capturing musical features without incurring excessive costs.

The effectiveness of transformer models in music similarity tasks depends on preprocessing choices, particularly the audio sample rate. Higher sample rates capture more musical detail

but increase computational costs, while lower rates reduce fidelity. Music spans a broad frequency spectrum, requiring a balance between preserving key features like timbre and harmony and maintaining efficiency. Research shows that a 22.05 kHz sample rate provides similar performance to 44.1 kHz while significantly reducing computational demands. This makes it a practical choice for transformer-based music similarity models.

C. Experimental adjustments for music similarity tasks

For Wav2Vec 2.0, the embedding size is set to 768, with 12 transformer layers, a learning rate of 0.0001, and a batch size of 32, using the Adam optimizer. CLAP utilizes a 512 embedding size, 12 layers, a learning rate of 0.0001, and a batch size of 64, designed for handling multimodal inputs. AST is configured with a 768 embedding size, 12 layers, a learning rate of 0.00005, and a batch size of 32, focusing on spectrogram-based audio representations. Mulan, which processes both audio and text, has a 512 embedding size, 6 transformer layers, a learning rate of 0.0001, and a batch size of 32. Each model includes a dropout rate of 0.1 to prevent overfitting, and the sequence lengths for the models are set at 500 spectrogram frames for AST and 512 tokens for CLAP and Mulan. These models are trained and evaluated on datasets including MSD, GZTAN, FSD50K, and Free Music Archive, which provide diverse music content for evaluating performance across different genres and formats.

During the experimentation process, several strategies were applied to optimize the performance of the models. First, the learning rates for each model were fine-tuned: for Wav2Vec 2.0, a lower learning rate of 0.00005 was used to prevent overshooting during fine-tuning, given its deep architecture and large pre-trained weights; for AST, learning rates of 0.0001 and 0.0002 were tested to achieve faster convergence without sacrificing accuracy, considering its focus on spectrograms; and for Mulan, a smaller learning rate of 0.00005 was chosen to improve the integration of both audio and text features. Additionally, learning rate schedulers, were implemented to adjust the learning rate dynamically during training to enhance convergence. To increase model robustness, data augmentation techniques like pitch shifting, time-stretching, and noise injection were applied for Wav2Vec 2.0 and AST. Pre-trained models for CLAP and Mulan were fine-tuned on music-specific datasets, such as FSD50K, to improve model accuracy by adapting the models to domain-specific music features.

D. Evaluation

The performance of the music similarity and recommendation system is assessed using multiple evaluation metrics to ensure a comprehensive analysis of its effectiveness. These metrics help quantify the relevance, ranking quality, and similarity of the recommended songs in relation to a given query song. Below, we describe each metric in detail, followed by a presentation of example evaluation tables illustrating the system's performance across different experimental conditions.

1) Evaluation Metrics: precision and recall [52] are fundamental metrics for evaluating the quality of music recommendations. Precision measures the proportion of relevant recommended songs, while recall assesses the proportion of relevant songs successfully retrieved by the system. However, precision and recall often have a trade-off, where increasing one may decrease the other. The F1-score balances by calculating their harmonic mean [53]. A higher F1 score indicates better overall performance. For example, if precision is 60% and recall is 40%, the F1-score provides a single metric 48% that reflects the balance between them.

Mean Average Precision [54] is a ranking-based metric that evaluates how well the recommendation system orders relevant songs in a ranked list. It computes the average precision for each relevant item in the list and then averages over all queries.

2) Results: present a comprehensive analysis of the transformer-based model evaluation, examining their performance in both standalone configurations and an integrated hybrid approach. By comparing individual models that process audio and text separately with a combined multimodal architecture, I assess the impact of feature fusion on music similarity and recommendation accuracy. The results highlight how leveraging multiple modalities enhances the system's ability to capture intricate relationships between songs, ultimately improving recommendation quality and interpretability.

TABLE II. PRECISION, RECALL AND F1-SCORE MODEL EVALUATION

Model	Modality	Precision	Recall	F1-score	MAP10
Wav2Vec2	Audio	0.72	0.65	0.7	0.76
CLAP	Audio	0.75	0.68	0.73	0.79
AST	Text	0.68	0.61	0.67	0.72
MuLan	Audio+Text	0.79	0.74	0.77	0.81
Hybrid	Audio+Text	0.87	0.82	0.85	0.88

The hybrid model combines both audio and text features to improve key aspects like precision, recall, and ranking accuracy. By structuring music similarity relationships within a semantic framework, the ontology enhances the interpretability of the recommendations. The use of cosine similarity helps to identify that songs within the same genre or by the same artist tend to have higher similarity scores, which aligns with how humans generally perceive music similarity. When comparing Wav2Vec2 and AST, the AST model performs slightly better in audio-only scenarios. This is because AST processes

spectrograms, capturing more detailed frequency and temporal patterns in the music. The Hybrid Model (Wav2Vec2 + AST + BERT), which combines Wav2Vec2, AST, and BERT, shows the best overall performance. This emphasizes the benefits of incorporating both audio (through Wav2Vec2 and AST) and text (through BERT) features to improve music similarity. By combining these different modalities, the model is able to extract richer, more comprehensive information, leading to more accurate predictions. When comparing MuLan with the Hybrid Model, the addition of Wav2Vec2 and AST in the hybrid approach enhances the feature extraction process, yielding better performance across all metrics. These results demonstrate that hybrid models, leveraging transformer-based architectures across various modalities, can significantly improve music similarity evaluation and provide more accurate, contextually relevant recommendations. MAP10 is a metric used to assess the quality of a recommendation system by focusing on the relevance of the top 10 recommendations.

The proposed transformer-based approach achieves strong performance but requires significant computational resources, which may limit deployment on smaller or resource-constrained platforms. To address this, the system can be adapted using lightweight transformer variants or lower-dimensional embeddings. Additionally, precomputing embeddings and using efficient retrieval methods can reduce real-time costs, improving scalability in practical settings.

While the study acknowledges the cold-start problem, the transformer-based approach mitigates it more effectively than traditional methods by leveraging pre-trained models and semantic embeddings. The model can infer similarity based on content alone by incorporating rich contextual information from audio, lyrics, and metadata, even for items with limited interaction history. This allows the system to recommend new or obscure songs without relying solely on user behavior data, offering a more robust solution to cold-start scenarios.

D. Building ontologies

Building an ontology from the evaluation of a music recommendation system involves structuring the key concepts and relationships between them in a formalized way.

1) Classes: based on the evaluation, the following classes could be relevant:

- Song: represents individual tracks of music.
- Artist: represents the creators or performers of songs.
- Genre: represents different categories of music.
- AudioFeature: Represents audio-based features of a song, such as rhythm, pitch, timbre, or beat.
- TextFeature: Represents text-based features of a song, such as lyrics or metadata.

2) Relationships: between the identified classes. These relationships will allow the ontology to represent how the entities are connected:

- hasArtist: connects a song to an artist.
- hasGenre: connects a song to a genre.
- hasAudioFeature: connects a song to its audio.
- hasTextFeature: Connects a song to its text features.
- similarTo: represents similarity between two songs based on shared attributes.

The structure of the proposed music ontology is illustrated in Fig. 2.

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:music="http://rumimi7157.jyu.fi/music#"
         xmlns:owl="http://www.w3.org/2002/07/owl#">

    <!--  Define Classes  -->
    <rdf:Class rdf:about="music:Song"/>
    <rdf:Class rdf:about="music:Artist"/>
    <rdf:Class rdf:about="music:Genre"/>
    <rdf:Class rdf:about="music:Mood"/>
    <rdf:Class rdf:about="music:AudioFeature"/>
    <rdf:Class rdf:about="music:TextFeature"/>
    <rdf:Class rdf:about="music:EvaluationMetric"/>
    <!--  Define Properties  -->
    <rdf:Property rdf:about="music:hasArtist"/>
    <rdf:Property rdf:about="music:hasGenre"/>
    <rdf:Property rdf:about="music:hasMood"/>
    <rdf:Property rdf:about="music:hasAudioFeature"/>
    <rdf:Property rdf:about="music:hasTextFeature"/>
    <rdf:Property rdf:about="music:hasPrecision"/>
    <rdf:Property rdf:about="music:similarTo">
        <rdf:type rdf:resource="owl:SymmetricProperty"/>
    </rdf:Property>
    <!--  Define Instances  -->
    <rdf:Description rdf:about="music:SongA">
    <rdf:type rdf:resource="music:Song"/>
    <music:hasArtist rdf:resource="music:ArtistX"/>
    <music:hasGenre rdf:resource="music:Rock"/>
    <music:hasMood rdf:resource="music:Energetic"/>
    </rdf:Description>
</rdf:RDF>
```

Fig. 2. Red, blue and green line

3) Evaluation Metrics: to reflect the evaluation the ontology includes evaluation metrics. Metrics can be modeled and linked to the Song class to show how well a recommendation matches the expected relevance, for example: Song A hasPrecision 0.8.

As the system evolves and new data or features are incorporated, the ontology should be updated to include new classes, relationships, and instances. This iterative process ensures that the ontology stays relevant and accurately represents the music domain. Integrating OWL-based ontologies enhances semantic understanding by structuring metadata and contextual relationships among musical entities. These ontologies are mapped to an embedding space through ontology-aware feature encoders, aligning symbolic descriptors with learned representations. During inference, this semantic layer aids in real-time filtering, disambiguation, and contextual enrichment of recommendations.

V. CONCLUSION

This study presents a transformer-based multimodal framework for music similarity analysis and recommendation, addressing ongoing personalization, diversity, and contextual relevance challenges. By independently extracting features from audio, lyrics, and metadata using state-of-the-art models such as Wav2vec, CLAP, AST, and Mulan and fusing them within a transformer architecture, the proposed system captures complex inter-modal relationships to generate more accurate and context-aware recommendations.

The novelty of our approach lies in its unified transformer-driven design, which enables deeper semantic integration across

modalities compared to traditional architectures. This leads to a more comprehensive understanding of music similarity, allowing the model to generalize across genres, cultural contexts, and listener behaviors. A modular multimodal processing pipeline may incorporate specialized encoders tailored for distinct data types. Additionally, a transformer-based fusion mechanism can be employed to capture high-level interrelationships across various modalities. Furthermore, a framework for integrating user interaction data will enhance implicit similarity modeling and support personalized experiences.

Empirical evaluations on benchmark datasets demonstrate the effectiveness of the proposed framework in improving both the accuracy and contextual relevance of music recommendations.

Each modality is processed independently to extract the relevant features before being integrated into a unified model. The feature extraction process ensures that each modality contributes its unique perspective on the music, capturing a holistic view of similarity. Afterward, these diverse features are merged within the transformer model, which is designed to learn complex relationships between the various input sources. This integrated approach allows the model to generate highly accurate and context-aware music recommendations, providing a comprehensive understanding of music similarity across different levels of analysis.

Several directions remain open for further research. First, incorporating temporal modeling of user behavior through sequence-based or recurrent approaches could enhance the system's responsiveness to evolving preferences. Second, exploring cross-lingual and culturally adaptive models would improve global applicability. Third, extending the framework with contrastive learning or self-supervised objectives may lead to more robust representations, especially in low-resource or cold-start scenarios. Finally, real-time deployment challenges must be addressed to bring such multimodal systems into production-grade environments, including scalability, latency, and privacy-aware design.

In conclusion, this study contributes to the growing body of research on deep learning-based music recommendation systems by demonstrating the effectiveness of transformer models in enhancing similarity analysis and personalization. Continued advancements in this area will be essential for developing more intelligent, context-aware, and inclusive music recommendation frameworks that cater to a global audience.

REFERENCES

[1] M. Velankar and P. Kulkarni, "Music Recommendation Systems: Overview and Challenges," in Advances in Speech and Music Technology, A. Biswas, E. Wennekes, A. Wieczorkowska, and R.H. Laskar, Eds., Signals and Communication Technology, 2023.

[2] D. Perera, M. Rajaratne, S. Arunathilake, K. Karunanayaka, and B. Liyanage, "A Critical Analysis of Music Recommendation Systems and New Perspectives", in Human Interaction, Emerging Technologies and Future Applications, T. Ahram, R. Taiar, V. Gremeaux-Bader, and K. Aminian, Eds., vol. 1152, Advances in Intelligent Systems and Computing, 2020.

[3] Y. Deldjoo, M. Schedl, and P. Knees, "Content-Driven Music Recommendation: Evolution, State of the Art, and Challenges", Comput. Sci. Rev., vol. 51, 2024.

[4] A. Klimashevskaia, D. Jannach Elahi, and M. Trattner, "A Survey on Popularity Bias in Recommender Systems", User Model. User-Adapt. Interact., 2024, doi: 10.1007/s11257-024-09406-0. Online publication date: 1-Jul-2024.

[5] S. Gupta, K. Kaur, and S. Jain, "EqBal-RS: Mitigating Popularity Bias in Recommender Systems", J. Intell. Inf. Syst., vol. 62, pp. 509–534, 2024.

[6] M. Waris, M. Zaman Fakhar, M. Gulsoy, E. Yalcin, and A. Bilge, "A Novel Pre-Processing Technique to Combat Popularity Bias in Personality-Aware Recommender Systems", IEEE Access, vol. 12, pp. 183230–183251, 2024.

[7] V.R. Revathy and S.P. Anitha, "Cold Start Problem in Social Recommender Systems: State-of-the-Art Review", in Advances in Computer Communication and Computational Sciences, S. Bhatia, S. Tiwari, K. Mishra, and M. Trivedi, Eds., vol. 759, Advances in Intelligent Systems and Computing, Springer, Singapore, 2019.

[8] D. Hesmondhalgh, R. Campos Valverde, D. Kaye, and Z. Li, "The Impact of Algorithmically Driven Recommendation Systems on Music Consumption and Production: A Literature Review", UK Centre for Data Ethics and Innovation Reports, Thousand Oaks, California, USA, 2023.

[9] D. Shakespeare, V. Chareyron, and C. Roth, "Reframing the Filter Bubble through Diverse Scale Effects in Online Music Consumption", Sci. Rep., Feb. 2025.

[10] M.A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-Based Music Information Retrieval: Current Directions and Future Challenges", Proc. IEEE, vol. 96, no. 4, pp. 668–696, Apr. 2008.

[11] C. Anuradha, R. Abhinaba, and H. Dorien, "MIRFLEX: Music Information Retrieval Feature Library for Extraction", Nov. 2024.

[12] Y.M.G. Costa, L.S. Oliveira, and C.N. Silla, "An Evaluation of Convolutional Neural Networks for Music Classification Using Spectrograms", Appl. Soft Comput., vol. 52, 2017.

[13] M. Bevec, M. Tkalčič, and M. Pesek, "Hybrid Music Recommendation with Graph Neural Networks", User Model User-Adap. Inter., vol. 34, pp. 1891–1928, 2024.

[14] M. Quadrana, P. Cremonesi, and D. Jannach, "Sequence-Aware Recommender Systems", 2019.

[15] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep Learning Based Recommender System: A Survey and New Perspectives", ACM Comput. Surv., vol. 50, no. 1, 2019

[16] X. Gu et al., "EEG-Based Brain-Computer Interfaces (BCIs): A Survey of Recent Studies on Signal Sensing Technologies and Computational Intelligence Approaches and Their Applications", IEEE/ACM Trans. Comput. Biol. Bioinformatics, vol. 18, no. 5, pp. 1645–1666, Sept.-Oct. 2021.

[17] O. Aruna, B. Venkata, and S. Naik, "Emotion Based Music Player", 2024.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need", in Proc. 31st Conf. Neural Inf. Process. Syst. (NIPS), Long Beach, CA, USA, Dec. 4–9, 2017.

[19] C.-Z.A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, C. Hawthorne, A.M. Dai, M.D. Hoffman, and D. Eck, "Music Transformer: Generating Music with Long-Term Structure", in Proc. Int. Conf. Learn. Represent. (ICLR), 2019.

[20] J. Thickstun, "The Transformer Model in Equations", University of Washington, Seattle, WA, 2021.

[21] Y. Lai, "A Comparison of Traditional Machine Learning and Deep Learning in Image Recognition", J. Phys. Conf. Ser., 2019.

[22] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "A Framework for Self-Supervised Learning of Speech Representations", 2020.

[23] Meta AI official website. Web: https://ai.meta.com

[24] Y. Gong, Y. Chung, and J. Glass, "AST: Audio Spectrogram Transformer", Interspeech, 2021.

[25] Q. Huang, A. Jansen, J. Lee, R. Ganti, J.Y. Li, and D.P.W. Ellis, "MuLan: A Joint Embedding of Music Audio and Natural Language", 2022.

[26] Y. Cai, Y. Liu, Z. Zhang, and J.Q. Shi, "CLAP: Isolating Content from Style through Contrastive Learning with Augmented Prompts", 2023.

[27] W. Huang, A. Wu, Y. Yang, X. Luo, Y. Yang, L. Hu, Q. Dai, X. Dai, D. Chen, C. Luo, and L. Qiu, "LLM2CLIP: Powerful Language Model Unlocks Richer Visual Representation", 2024.
[28] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows", Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021.
[29] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach", 2019.
[30] M. Zeng, X. Tan, R. Wang, Z. Ju, T. Qin, and T.-Y. Liu, "MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training", 2021.
[31] R. Jahangir, Y.W. Teh, H.F. Nweke, G. Mujtaba, M.A. Al-Garadi, and I. Ali, "Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges", Expert Syst. Appl., vol. 171, 2021.
[32] MFCC implementation and tutorial, Kaggle website, Web: https://www.kaggle.com/code/ilyamich/mfcc-implementation-and-tutorial.
[33] An audio event embedding model vaggish Kaggle documentation, Web: https://www.kaggle.com/models/google/vggish.
[34] OpenL3 is an open-source Python library documentation, Web: https://openl3.readthedocs.io/en/latest.
[35] Natural language processing definition, Web: https://www.ibm.com/think/topics/natural-language-processing.
[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", 2018.
[37] OpenAi GPT official website, Web: https://openai.com/index/chatgpt
[38] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks", 2019.
[39] Spotify API official documentation, Web: https://developer.spotify.com/documentation/web-api
[40] Million Song Dataset official web page, Web: http://millionsongdataset.com.
[41] R. Shankar, K. Tan, B. Xu, and A. Kumar, "A Closer Look at Wav2Vec2 Embeddings for On-Device Single-Channel Speech Enhancement", 2024.
[42] X. Shi, X. Li, and T. Toda, "Multimodal Fusion of Music Theory-Inspired and Self-Supervised Representations for Improved Emotion Recognition", Interspeech, 2024.
[43] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional Feature Fusion", IEEE Winter Conf. Appl. Comput. Vis. (WACV), Waikoloa, HI, USA, Jan. 2021, pp. 3559–3568.
[44] J. Han, M. Kamber, and J. Pei, "Getting to Know Your Data", in Data Mining, 3rd ed., 2012.
[45] F.E. Szabo, "The Linear Algebra Survival Guide", Academic Press, 2015.
[46] H. Yin, G. Song, L. Zhang, and C. Wu, "Chapter 1 - Introduction: Virtual experiments with iBEM", Academic Press, 2022.
[47] D. Korzun, S. Yalovitsyna, and V. Volokhova, "Smart services as cultural and historical heritage information assistance for museum visitors and personnel", Balt. J. Mod. Comput., vol. 6, Dec. 2018.
[48] GTZAN Dataset - Music Genre Classification, Web: https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification
[49] D. E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," 2022.
[50] Creative commons official website, Web: https://arxiv.org/pdf/2010.00475v2
[51] Free Music Archive official website, Web: https://freemusicarchive.org/.
[52] Precision and Recall in Machine Learning definition, Web: https://www.analyticsvidhya.com/articles/precision-and-recall-in-machine-learning/.
[53] Inside Machine Learning, Recall, Precision, F1 Score – Simple Metric Explanation Machine Learning, Web: https://inside-machinelearning.com/en/recall-precision-f1-score-simple-metric-explanation-machine-learning/.
[54] Kili Technology, Mean Average Precision: A Complete Guide, Web: https://kili-technology.com/data-labeling/machine-learning/mean-average-precision-map-a-complete-guide.