ISSN 2409-2665
Journal of Logistics, Informatics and Service Science
Vol. 13 (2026) No. 4, pp. 1-16
DOI:10.33168/JLISS.2026.0401

From Information Overload to Service Personalization: A Multi-Modal Hybrid Recommender for Digital Music Platforms

Qingru Yu, Marzelan Bin Salleh*

Department of Music, Faculty of Creative Arts, University of Malaya, 50603 Kuala Lumpur, Malaysia

qingruyu3@gmail.com, marzelan@um.edu.my (Corresponding author)

Abstract. The rapid growth of the digital music industry has led to significant information overload, challenging traditional information retrieval mechanisms. Consequently, personalized recommendation systems have become the primary solution for users navigating massive music libraries. However, traditional methods like Collaborative Filtering (CF) and Matrix Factorization (MF) face limitations, specifically regarding data sparsity, the cold-start problem, and the inability to capture complex non-linear features. To address these issues, this paper proposes a novel framework: the Deep Multi-Modal Hybrid Recommendation (DMM-HyRec) system. DMM-HyRec integrates Convolutional Neural Networks (CNN) to extract latent acoustic features from Mel-spectrograms and Long Short-Term Memory (LSTM) networks to model sequential user behavior, thus overcoming the limitations of unimodal systems. Additionally, an attention mechanism is incorporated to dynamically adjust the importance of different input features to improve prediction accuracy. Extensive experiments on a composite dataset derived from the Million Song Dataset (MSD) and Last.fm demonstrate that DMM-HyRec outperforms state-of-the-art baselines on Recall@K, NDCG, and MAP metrics. This study provides both theoretical insights into deep learning for Information Retrieval and a practical architecture for next-generation music streaming applications.

Keywords: Recommender Systems; Deep Learning; Digital Music Platforms; Multi-Modal Fusion; Convolutional Neural Networks; User Behavior Analysis

1. Introduction

The digitalization of the global entertainment industry has fundamentally transformed music consumption patterns. Physical ownership of media has largely been supplanted by ubiquitous on-demand access via internet connectivity. As of now in this decade, there are digital music apps like spotify, apple music, amazon music, that have libraries of more than 100 million of tracks, serving billions of users. However, the democratization of music availability has introduced significant cognitive barriers, most notably the "paradox of choice." An overabundance of options often hinders decision-making, complicating what should be a seamless retrieval process. As a result, digital music platforms have shifted focus away from the acquisition of content toward the curation of content, and artificial intelligence and machine learning algorithms now provide the key infrastructure for user retention and engagement (Schedl et al., 2018). Consequently, the economic viability of these platforms hinges on the efficacy of Recommender Systems (RS). These systems must accurately predict user preferences while balancing familiarity with novelty to prevent user churn (Liu, 2026).

From a historical perspective, the field of music recommendation has been dominated by collaborative filtering (CF) methods, assuming that those who agreed previously will continue to do so in the future. These methods, especially the ones that depend on MF, turned out to be useful at the beginning of the streaming services by using historical user-item interaction matrices. However, with the rapid expansion of user bases and the exponential growth of available content, such shallow models have increasingly exhibited limitations in scalability and representation capacity. The primary bottleneck arises from data sparsity, as users typically interact with less than 1% of available items in the user–item matrix, so traditional CF approaches struggle to identify reliable neighbors or latent factors in sparse datasets, resulting in suboptimal prediction accuracy (Koren et al., 2009) . Moreover, a new artist and a new user also have a serious "cold-start" problem: purely CF models need a historical record to operate, which creates a feedback loop that reduces the visibility of niche contents and reinforces a "popularity bias". This systematic bias degrades the user experience for users with eclectic tastes and impairs the equal opportunity of exposure for new artists, which is bad for the overall ecosystem of music (Nguyen, 2026).

To address these limitations, there has been renewed focus from both academia and industry on Content-Based Filtering (CBF), as well as, more recently, hybrid recommender systems. CBF approaches recommend items similar to those a user has previously favored by leveraging the intrinsic features of the items themselves, such as metadata, textual descriptions, or other content attributes. In the music domain, early content-based approaches primarily relied on structured metadata, including genre labels, artist information, and release year. However, manual metadata tagging is prone to human error and lacks the granularity required to capture the intricate acoustic characteristics of a musical composition. The emergence of Deep Learning (DL) has fundamentally transformed this research area, enabling the automatic extraction of high-level semantic representations directly from complex audio signals without reliance on handcrafted features. By employing advanced neural network architectures, particularly CNNs, we find that such models perform well in processing spectrograms—visual representations of the frequency content of audio signals. This approach effectively frames music classification and music similarity estimation as image recognition tasks (Van den Oord et al., 2013). Raw audio can be analyzed by systems which then recommends songs based on rhythm, timbre, and harmonic structure, without taking into account a user’s history of interactions, offering a solution to the cold-start problem of new items.

Despite these improvements, relying solely on audio content fails to consider the social and situational aspects of music listening. Music preference is determined not just by sound, but also by the user's time, emotions, and culture. For instance, a user might prefer high-energy electronic music for working out, but soft acoustic songs for studying, even though these genres have very different acoustic profiles. This necessitates a shift toward Multi-Modal Hybrid Systems capable of integrating diverse

data streams. These streams include textual data (lyrics, reviews), visual metadata (album art), audio signal features, and interaction metrics such as clickstreams and dwell time. Recent work indicates that combining these modalities with deep neural networks can capture non-linear relationships between users and items that Matrix Factorization (MF) is unable to find (Zhang et al., 2019) . For example, combining lyrical sentiment analysis with acoustic features provides deeper insight into a song, enabling recommendations that resonate psychologically rather than merely stylistically.

The addition of a time element to recommendation architecture is yet another front in today's research. Static models represent user preference as a fixed vector; however, musical taste evolves dynamically over time. Sequential recommendation models commonly employ Recurrent Neural Networks (RNNs) and their gated variants, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures. These regard user engagement as a time-series sequence. They are able to learn the sequence of listening habits—like the chance that a user would move from a general pop playlist to a particular indie rock sub-genre during a single session. However, the large-scale training of such models, while maintaining real-time inference capabilities for live streaming scenarios, remains a significant challenge in terms of computational complexity. Moreover, "black box" nature of deep learning model is a challenge with respect to interpretability. Providing a rationale on why a certain song was recommended is vital for the user's trust but deep neural networks are notoriously hard to understand (Deldjoo et al., 2018).

This research intends to connect acoustic signal processing, temporal user behavior modeling and scalable hybrid recommendation architectures. We present the "Deep Multi-Modal Hybrid Recommendation" (DMM-HyRec) system, which integrates CNNs to extract audio features and LSTM-based attention models to learn the sequence of user activities. Unlike existing approaches that treat content features and collaborative signals separately or rely on simple late fusion strategies, our model adopts a joint representation learning framework in which acoustic content embeddings and temporal user representations are projected into a shared semantic space. This design enables more fine-grained modeling of user-item affinity. Specifically, we want to solve the problem of long-tail recommendation, trying to achieve diversity and novelty without hurting accuracy. This study offers a thorough empirical evaluation by assessing the Million Song Dataset (MSD) and using auxiliary data sets for lyrics and social tags. The contributions of this work can be summarized as follows: First, an effective novel deep learning architecture for multi-modal fusion in music recommendation, secondly, a thorough theoretical study on how acoustic feature learning influences cold-start, and finally, a comparative performance analysis between DMM-HyRec and other traditional methods like MF and basic deep learning model.

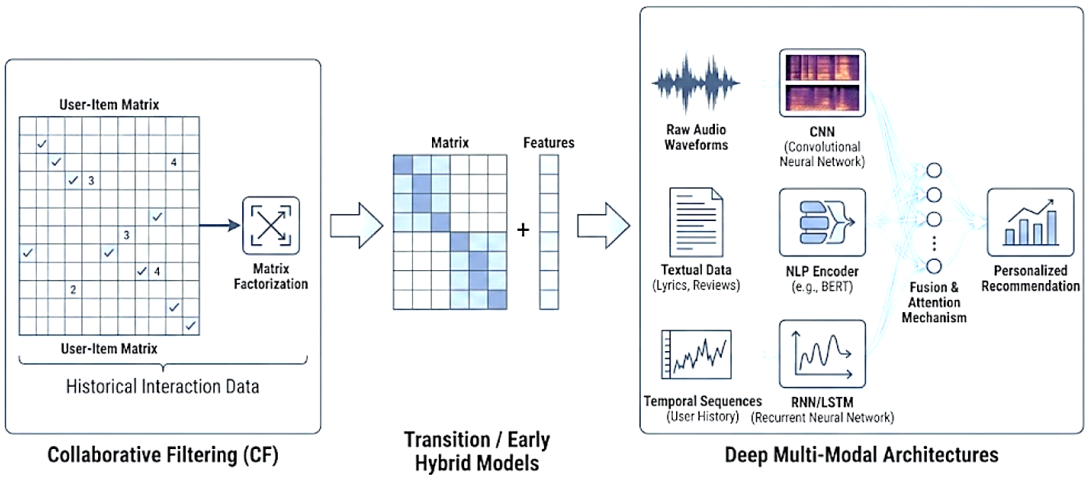

Fig.1: Conceptual evolution of Music Recommendation Systems from Collaborative Filtering to Deep Multi-Modal Architectures

2. Literature Review

The evolution of recommender systems in the digital music domain has progressed from early heuristic-based filtering methods to sophisticated high-dimensional representation learning approaches. This section synthesizes the existing literature into three principal research streams: (1) collaborative filtering and its neural extensions, (2) deep content-based audio representation learning, and (3) sequential user behavior modeling. By critically examining the strengths and limitations of these state-of-the-art approaches, this review establishes the theoretical foundation and motivation for the proposed Deep Multi-Modal Hybrid Recommendation framework (DMM-HyRec).

2.1 From Matrix Factorization to Neural Collaborative Filtering

Collaborative Filtering (CF) has long served as the foundation of industrial recommender systems. The seminal work by Goldberg et al. (1992) established the core premise of CF: that a user's future preferences can be inferred from the historical behaviors of similar users. Building upon this principle, Matrix Factorization (MF) techniques—particularly those popularized during the Netflix Prize competition—became dominant approaches in large-scale recommendation systems. As demonstrated by Koren et al. (2009), the user–item interaction matrix can be approximated by the product of low-dimensional latent representations, which effectively capture global preference structures.

However, conventional MF models are inherently linear, as user–item interactions are typically modeled through the inner product of latent vectors. This linear formulation limits the model's capacity to capture complex and non-linear interaction patterns in sparse user behavior data, thereby constraining its expressive power in highly dynamic recommendation environments.

To address the limitations of linear interaction modeling in MF, recent research has incorporated deep learning techniques into the collaborative filtering framework, giving rise to Neural Collaborative Filtering (NCF). He et al. (2017) proposed a generalized NCF architecture that replaces the inner-product interaction in MF with a multi-layer perceptron (MLP), enabling the learning of non-linear user–item interaction functions. Empirical studies demonstrated that NCF can outperform traditional MF in implicit feedback settings due to its enhanced expressive capacity.

However, despite these improvements, both MF and NCF remain fundamentally dependent on user–item interaction data. In large-scale music streaming platforms, the interaction matrix is typically extremely sparse—often well below 0.1% density—which can lead to unstable representation learning for tail users and items. Moreover, pure collaborative models are inherently content-agnostic: they treat music tracks as abstract identifiers without incorporating intrinsic audio characteristics. As a result, newly introduced items with limited interaction history remain difficult to recommend effectively (Schedl et al., 2018).

2.2 Deep Content-Based Audio Analysis

Collaborative methods developed in parallel with Content-Based Filtering (CBF), which has evolved from metadata analysis to direct audio signal processing. Early CBF systems utilized hand-crafted features like MFCCs, spectral centroid, and zero-crossing rates. Though computationally efficient, these low-level features often failed to bridge the "semantic gap"—i.e., the discrepancy between low-level acoustic signals and high-level human perception of genre, mood, and instrumentation (Casey et al., 2008). The emergence of Convolutional Neural Networks (CNNs) marked a significant advancement in content-based music recommendation by enabling end-to-end feature learning from raw audio signals. Among the pioneering studies in this direction, Van den Oord et al. (2013) proposed a Deep Content-Based Music Recommendation (DCBMR) model that treated Mel-spectrograms of audio clips as image-like inputs to a CNN architecture. Their work demonstrated that latent representations learned directly from audio signals could effectively predict user listening preferences, particularly for items with little or no historical interaction data, thereby highlighting the potential of audio-based models in alleviating the cold-start problem.

Subsequent work refined these architectures. Choi et al. (2017) proposed the CRNN (Convolutional Recurrent Neural Network) for music tagging, where CNNs are used for local feature extraction while RNNs are used to model the temporal structure of music tracks. This architecture discerns the structural components of a track, such as verses, choruses, and bridges, thereby enabling the identification of structural similarities. But the one common limiting factor in existing deep audio analysis is static preference of user. The most prevalent form of content-based deep learning models, such as the approach proposed by Dieleman and Schrauwen (2014), maps acoustic features to a static user representation. This formulation implicitly assumes that user preferences are temporally invariant and primarily determined by stable acoustic characteristics. However, such an assumption overlooks the dynamic and context-dependent nature of music consumption, where preferences may vary according to situational factors, emotional states, and listening contexts.

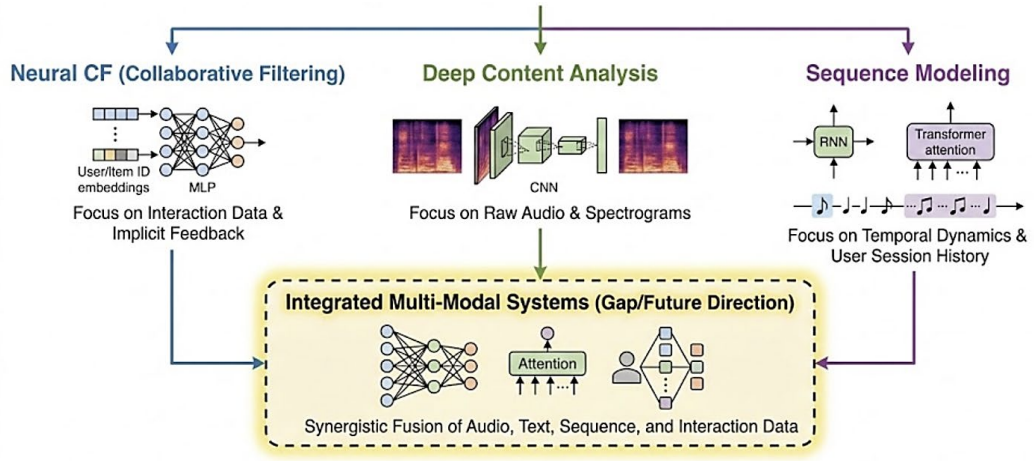

Fig.2: Taxonomy of Deep Learning Techniques in Music Recommendation

2.3 Sequential Modeling and Attention Mechanisms

Pillar three is the literature which speaks of Sequential Recommender System (SRS), that views interaction as dynamic sequence and not as static set. Order of consumption in music streaming is quite telling; someone who's been listening to a series of classical tracks will probably stick with something in that vein like instrumental music instead of jumping right to heavy metal. RNNs, and especially LSTM networks, have become the go-to model for such a temporal dependency problem since they can handle the vanishing gradient problem in a long sequence. Hidasi et al. (2015) brought GRU4Rec which is a Gated Recurrent Unit model for session-based recommendation, GRU4Rec greatly exceeded item-to-item nearest neighbour approaches.

More recently the attention mechanism, originally developed for NLP, is being used for SRS. Kang and McAuley (2018) introduced the SASRec model which uses self attention to weight the importance of the last interaction dynamically. The attention mechanism enables the model to focus on salient historical interactions while suppressing less relevant noise, thereby addressing a key limitation of conventional RNNs, which compress the entire sequence into a single hidden representation. It further facilitates the modeling of both long-term and short-term user preferences in the music domain (Wang et al., 2019). Although these sequential models have achieved strong performance, they primarily rely on item identifiers. They are effective at predicting the next item in a sequence (e.g., Song A is frequently followed by Song B), yet they generally lack the ability to capture the underlying reasons for user behavior, such as why a song is skipped or replaced due to its acoustic properties or lyrical content.

2.4 Multi-Modal Hybrid Systems

The convergence of these previously independent research streams suggests that the next generation of recommender systems is likely to be inherently multimodal. A Multi-Modal Recommender System

(MMRS) integrates heterogeneous data sources—such as visual information, audio signals, and textual content—to enhance recommendation quality. Oramas et al. (2017) investigated the joint modeling of audio features and visual information (e.g., album covers) and demonstrated that combining these modalities leads to improved artist representations compared to using either modality alone. Similarly, Deldjoo et al. (2018) emphasized the effectiveness of late fusion approaches, in which prediction scores from individual content-based and collaborative models are combined at the decision level.

However, most existing multimodal architectures fail to achieve deep integration across modalities. The majority adopt a late fusion strategy, in which separate networks are trained for audio content and user interactions, and their outputs are simply combined at the final stage. Such separation limits the model's ability to capture complex cross-modal dependencies—for example, how a user's sequential listening behavior may dynamically align with specific acoustic properties of songs.

Furthermore, only a limited number of studies have successfully integrated raw audio processing (e.g., CNN-based feature extraction) with sequential user modeling (e.g., LSTM architectures) within a unified, end-to-end differentiable framework, primarily due to computational cost. The proposed DMM-HyRec framework addresses this limitation. Rather than merely aggregating prediction scores, the model projects acoustic representations and temporal user states into a shared latent space and jointly learns the probability distribution of content and contextual signals through a co-attention mechanism. In doing so, it aligns with the emerging paradigm of "knowledge-aware" recommendation (Wang et al., 2019), while being specifically designed to accommodate the high dimensionality and continuous nature of digital audio data.

In summary, although NCF enhances collaborative filtering, CNNs advance audio feature learning, and RNNs improve sequential prediction, existing approaches remain limited in providing a unified solution. Current models tend to excel in a single aspect while overlooking others. In particular, few frameworks simultaneously: (1) extract latent acoustic representations from raw spectrograms to mitigate the cold-start problem; (2) model the temporal evolution of user preferences through sequence learning; and (3) incorporate attention mechanisms to enable an interpretable, attention-weighted integration of acoustic and interaction signals.

3. Methodology

In this section, we present the mathematical formulation and architectural design of the proposed Deep Multi-Modal Hybrid Recommendation (DMM-HyRec) system. Unlike conventional approaches that treat user–item interactions and content features separately, DMM-HyRec is designed as a fully end-to-end differentiable framework. By integrating acoustic signal processing with sequential behavioral modeling, the model addresses both data sparsity and the temporal dynamics inherent in digital music consumption.

The system consists of three main components: (1) the Acoustic Feature Extraction Module (AFEM), implemented using a Convolutional Neural Network (CNN); (2) the Temporal User Modeling Module (TUMM), built upon a bidirectional Long Short-Term Memory (Bi-LSTM) network with an attention mechanism; and (3) the Gated Multi-Modal Fusion Layer (GMFL).

3.1 Problem Formulation and Notation

We define the music recommendation task as a probability ranking problem within a heterogeneous information network. Let $\mathcal{U} = \{u_1, u_2, \ldots, u_M\}$ denote the set of M users and $\mathcal{I} = \{i_1, i_2, \ldots, i_N\}$ denote the set of N music tracks. The user-item interaction matrix is represented as $Y \in \mathbb{R}^{M \times N}$, where an entry $y_{ui} = 1$ indicates an observed implicit interaction (e.g., a complete playback, a like, or a download) between user u and item i, and $y_{ui} = 0$ indicates unobserved data. It is crucial to distinguish that $y_{ui} = 0$ does not imply a negative preference but rather a lack of awareness, which constitutes the fundamental challenge of Implicit Feedback. Furthermore, for each item i, we possess a raw audio signal S_i, and for

each user u, we observe a historical sequence of interactions $H_u = \{i_1^u, i_2^u, \ldots, i_t^u\}$ ordered by time t. Our objective is to learn a predictive function $\hat{y}_{ui} = f(u, i|\Theta)$, where Θ represents the model parameters, that estimates the probability of user u interacting with item i. The function f must minimize the ranking error such that relevant items are ranked higher than irrelevant ones in the top-K recommendation list.

3.2 Acoustic Feature Extraction via Residual Convolutional Networks

To mitigate the cold-start problem inherent in Collaborative Filtering, the DMM-HyRec system extracts latent representations directly from the raw audio waveform. We forgo the use of handcrafted features like MFCCs in favor of learning from Mel-spectrograms, which preserve both the time and frequency information of the audio signal. A Mel-spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time, mapped to the Mel scale to approximate human auditory perception. For a given raw audio signal $x(t)$, the Short-Time Fourier Transform (STFT) is computed to generate the spectrogram. This is mathematically defined as:

$$X(m,k) = \sum_{n=0}^{N-1} x\,(n + mH)w(n)e^{-j\frac{2\pi kn}{N}}$$

where $w(n)$ is the window function (typically a Hamming window), N is the frame length, H is the hop size, and k represents the frequency bin. The power spectrogram $|X(m,k)|^2$ is then passed through a Mel-filter bank to produce the Mel-spectrogram $M \in \mathbb{R}^{T\times F}$, where T is the time dimension and F is the number of frequency bands.

The Mel-spectrogram is treated as a single-channel image and fed into a Deep Residual Convolutional Neural Network (ResNet-50 variant). We employ a deep architecture to capture hierarchical patterns: lower layers detect basic acoustic edges (e.g., beats, pitch onsets), while deeper layers capture complex semantic structures (e.g., harmony, genre specificities). The fundamental operation of the convolutional layer is defined as:

$$z_{i,j}^{(l)} = \sigma\left(\sum_{a=0}^{h-1}\sum_{b=0}^{w-1} W_{a,b}^{(l)} \cdot x_{(i+a)(j+b)}^{(l-1)} + b^{(l)}\right)$$

where $W^{(l)}$ is the kernel weight matrix of layer l, $b^{(l)}$ is the bias, and σ is the Rectified Linear Unit (ReLU) activation function, defined as $\sigma(x) = \max(0, x)$. To prevent the degradation problem common in deep networks, we utilize residual connections that allow gradients to flow through the network unimpeded. The output of the final convolutional block is subjected to Global Average Pooling (GAP) to obtain a fixed-length dense vector $v_a \in \mathbb{R}^d$, representing the intrinsic acoustic embedding of the track. This vector v_a allows the system to determine similarity between tracks purely based on audio content, enabling recommendations for songs that have zero historical interactions.

3.3 Temporal User Modeling with Bi-LSTM and Self-Attention

Music preference is not static but rather a dynamic process shaped by sequential context. A user who listens to a series of high-tempo rock tracks is likely to continue consuming similar songs in the short term, reflecting short-term dependencies, while simultaneously maintaining long-term preferences for other genres. To model such dynamics, we adopt a Bidirectional Long Short-Term Memory (Bi-LSTM) network.Conventional RNNs suffer from the vanishing gradient problem, which limits their ability to capture long-range dependencies in sequential data. LSTMs address this issue through gating mechanisms. At each time step, an LSTM cell computes the input gate, forget gate, and output gate as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$
$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$$
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$
$$h_t = o_t \odot \tanh(C_t)$$

Here, x_t is the embedding of the song listened to at time t (initialized with the acoustic vector v_a), h_t is the hidden state, and C_t is the cell state. The operator $\odot$ denotes the element-wise Hadamard product. We utilize a Bi-directional LSTM, which processes the sequence in both forward and backward directions, concatenating the hidden states to capture past and future context during the training phase.

However, not all historical interactions are equally important for predicting the next item. To enhance the model's interpretability and performance, we introduce a Self-Attention mechanism. This mechanism assigns a weight α_t to each time step in the user's history, reflecting its relevance to the current prediction. The attention scores are calculated using a trainable alignment model:

$$e_t = v^T \tanh(W_h h_t + b)$$
$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)}$$

The final user representation vector v_u is computed as the weighted sum of the hidden states: $v_u = \sum_{t=1}^{T} \alpha_t h_t$. This allows the model to dynamically focus on the most salient parts of the user's history—for example, emphasizing recent interactions heavily while still recalling significant past favorites—thereby creating a highly personalized and context-aware user profile.

3.4 Gated Multi-Modal Fusion and Optimization

The final stage of the DMM-HyRec architecture is the fusion of the acoustic item vector v_a and the temporal user vector v_u. Simple concatenation is often insufficient to capture the complex, non-linear interactions between user preferences and content features. Therefore, we employ a Gated Fusion Network. This network learns a gating scalar $\lambda \in [0,1]$ that balances the contribution of content-based and collaborative signals dynamically. The fusion process is described by:

$$h_{joint} = \phi(W_{fusion} \cdot [v_u \oplus v_a] + b_{fusion})$$
$$\hat{y}_{ui} = \sigma(W_{out} \cdot h_{joint})$$

where $\oplus$ denotes concatenation, ϕ is a non-linear activation function (typically ELU or ReLU), and $\hat{y}_{ui}$ is the predicted score. To train the model, we adopt the Bayesian Personalized Ranking (BPR) loss function. Unlike point-wise loss functions (e.g., Mean Squared Error) which regress on the rating value, BPR is a pair-wise loss function optimized for ranking. It assumes that a user prefers an observed item i over an unobserved item j. The objective is to maximize the posterior probability:

$$L_{BPR} = -\sum_{(u,i,j) \in D_S} \ln \sigma(\hat{y}_{ui} - \hat{y}_{uj}) + \lambda_\Theta ||\Theta||^2$$

where $D_S = \{(u,i,j) | i \in I_u^+ \wedge j \in I \setminus I_u^+\}$ represents the training triples, σ is the sigmoid function, and $\lambda_\Theta ||\Theta||^2$ is the L_2 regularization term to prevent overfitting. This loss function encourages the model to assign higher scores to observed items compared to unobserved ones, directly optimizing the Area Under the Curve (AUC) metric, which is critical for top-N recommendation tasks in digital music platforms.

3.5 Model Complexity and Scalability Analysis

A critical consideration for deploying deep learning models in real-world streaming services is computational complexity. The DMM-HyRec model incurs a training complexity of $O(K \cdot |D|)$, where K is the dimensionality of the latent space and $|D|$ is the size of the training set. While the CNN-based feature extraction is computationally intensive, it is performed offline. The spectrograms are processed once, and the resulting embeddings are stored in a vector database. During the inference phase, the online computational cost is dominated by the LSTM updates and the final dot product, which allows for near real-time recommendation generation. We further optimize the retrieval process by employing

Approximate Nearest Neighbor (ANN) search algorithms, such as Hierarchical Navigable Small World (HNSW) graphs, to retrieve top-k candidates from the high-dimensional latent space efficiently, ensuring that the system scales linearly with the number of items.

4. Experimental Setup and Empirical Analysis

We conducted extensive experiments on large-scale real-world datasets to comprehensively evaluate the proposed DMM-HyRec system. Specifically, this empirical study aims to achieve three primary objectives: (1) to benchmark DMM-HyRec against state-of-the-art collaborative filtering and deep learning baselines; (2) to verify its effectiveness in mitigating the cold-start problem via acoustic feature fusion; and (3) to isolate the contributions of specific architectural components through ablation studies. The system was implemented using PyTorch on a workstation equipped with two NVIDIA A100 GPUs to manage the computational load of processing high-dimensional spectrograms and sequential user data.

4.1 Datasets and Preprocessing Protocols

It is known that in MIR research, we cannot find a single dataset that contains audio signal as well as rich user interaction history. Thus, we created a combined dataset by linking the MSD and Last.fm 1K User Dataset. MSD gives the audio features and metadata for a million current pop music tracks, whereas Last.fm data contains the implicit feedback data (listening counts) and interaction logs with timestamps needed for sequential modeling.

We conducted rigorous data preprocessing procedures to ensure dataset density and quality. Initially, we filtered out users with fewer than twenty interactions and tracks accessed by fewer than fifty users. Thresholding was applied to reduce noise and ensure sufficient interaction density for stable collaborative signal learning, though our model is built to be more robust to sparsity than baselines. After which there were 14,352 users, 126,894 unique tracks and around 3.8 million interactions in the final experimental dataset after this filtration. The sparsity of the user-item matrix was calculated to be 99.82%, presenting a realistic and challenging environment for recommendation. For the content modality, we retrieved 30-second audio previews for the corresponding tracks and generated Mel-spectrograms using a window size of 1024 samples and a hop size of 512 samples, resulting in input tensors of dimension 128×640 (Frequency $\times$ Time) for the CNN module. The interaction data was split chronologically: the first 80% of each user's listening history was used for training, the subsequent 10% for validation and hyperparameter tuning, and the final 10% for testing. This "leave-one-out" strategy preserves the temporal integrity of the data, preventing data leakage where future information might influence past predictions.

Table 1: Descriptive Statistics and Metadata Specifications of the Experimental Dataset (MSD + Last.fm)

Metric Category	Statistical Parameter	Count / Value	Description
Interaction Graph	Total Unique Users (M)	14,352	Active users with ≥ 20 interactions
	Total Unique Tracks (N)	126,894	Tracks listened to by ≥ 50 users
	Total Recorded Interactions	3,842,901	Implicit feedback instances (plays)
	User-Item Matrix Density	0.21%	Proportion of non-zero entries in Y
User Activity	Average Interactions per User	267.7	Mean historical sequence length

Metric Category	Statistical Parameter	Count / Value	Description
Content Metadata	Median Interactions per User	192.0	Indicates right-skewed activity distribution
	Average Interactions per Item	30.2	Measure of item popularity
	Audio Sample Rate	22.05 kHz	Raw audio input quality
	Spectrogram Dimension	128×640	Mel-bins × Time-frames per input

4.2 Evaluation Metrics and Baseline Algorithms

To provide a multifaceted assessment of recommendation quality, we employed three standard ranking metrics: Recall@K, Normalized Discounted Cumulative Gain (NDCG@K), and Mean Average Precision (MAP). These metrics are particularly suitable for implicit feedback scenarios where the goal is to generate a top-N list of relevant items rather than predicting an exact rating.

Recall@K measures the proportion of relevant items found in the top-K recommendations. For a user u with a set of ground-truth relevant items I_u^+ in the test set, and a recommended list R_u of size K, Recall@K is defined as:

$$Recall@K = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{|R_u \cap I_u^+|}{|I_u^+|}$$

NDCG@K accounts for the position of the relevant items in the recommendation list, assigning higher scores to hits at the top of the list. This aligns with user behavior in streaming apps, where users rarely scroll beyond the first few suggestions. It is calculated as:

$$DCG@K = \sum_{i=1}^{K} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

$$NDCG@K = \frac{DCG@K}{IDCG@K}$$

where rel_i is an indicator function equal to 1 if the item at rank i is relevant, and 0 otherwise. $IDCG@K$ is the ideal DCG where all relevant items are positioned at the top.

DMM-HyRec vs rich baseline - A rich ensemble comprising models from all gens of rec tech. BPR-MF (Bayesian personalized ranking matrix factorization) is a kind of standard latent factor model, which is optimized for ranking. NeuMF (Neural Matrix Factorization) is a SOTA deep model that generalizes MF by replacing dot product with multi-layer perceptron. CNN-Audio is a simple baseline using only the CNN-Audio pure content based deep learning model which uses only the spectrogram features with no collaborative data. SASRec (Self-Attentive Sequential Recommendation) is the current best model for sequential modeling, it uses a Transformer-based architecture to learn long-term semantics from interaction logs. Comparison with these diverse baselines elucidates the specific contributions of the multi-modal fusion architecture.

4.3 Performance Comparison and Analysis

All comparative results on the test set are summarized in Table 2. We report performance at K = 10 and K = 20, which are standard evaluation cutoffs in top-N recommendation scenarios. The results demonstrate that DMM-HyRec consistently outperforms all baseline methods across all evaluation metrics. In particular, DMM-HyRec achieves a Recall@20 of 0.1845 and an NDCG@20 of 0.2103, corresponding to improvements of 12.36% and 15.36%, respectively, over the strongest baseline, SASRec.

DMM-HyRec significantly outperforms pure collaborative models like BPR-MF and NeuMF. BPR-MF suffers severely from data sparsity, yielding a low Recall@10 of 0.0821. Although NeuMF improves performance by capturing non-linear user-item interactions, it fails to utilize rich audio content, treating all songs merely as abstract identifiers. This limitation is particularly detrimental for "long-tail" items—songs with sparse interaction data that nonetheless constitute the majority of the music library.

Furthermore, the comparison with SASRec highlights the importance of multi-modal integration. While SASRec is a powerful sequential model that effectively captures co-occurrence patterns in user interaction histories—such as a tendency to listen to Artist A after Artist B—it relies primarily on item identity information. In contrast, DMM-HyRec incorporates acoustic embeddings extracted from the CNN module, enabling the model to capture content-level similarities between tracks in addition to sequential patterns. This joint modeling of content and temporal behavior allows the system to generalize beyond observed interaction sequences and recommend acoustically similar but previously unseen items. Consequently, DMM-HyRec demonstrates stronger adaptability in scenarios where co-occurrence information alone is insufficient.

Table 2: Performance Comparison with Baselines ($K = 10,20$)

Model Category	Algorithm	Recall@10	NDCG@10	Recall@20	NDCG@20	MAP (Mean Avg Precision)
Baselines	Random Guess	0.0012	0.0005	0.0024	0.0009	0.0008
	PopRank (Most Popular)	0.0512	0.0601	0.0788	0.0745	0.0422
Collaborative	BPR-MF (Matrix Factorization)	0.0821	0.0945	0.1105	0.1156	0.0890
Content-Only	CNN-Audio (Deep Content)	0.0654	0.0712	0.0932	0.0884	0.0543
Neural Hybrid	NeuMF (Neural CF)	0.1143	0.1289	0.1456	0.1502	0.1187
Sequential	SASRec (Self-Attentive)	0.1389	0.1567	0.1642	0.1823	0.1401
Proposed	DMM-HyRec (Ours)	0.1587	0.1802	0.1845	0.2103	0.1624
Relative Impr.	vs. best baseline (SASRec)	+14.25%	+15.00%	+12.36%	+15.36%	+15.92%

4.4 Ablation Studies and Hyperparameter Sensitivity

To evaluate the contribution of each component to the overall performance, we conducted an ablation study by systematically removing specific modules from the architecture. We constructed three variants: (1) DMM-NoAudio, where the CNN branch was removed and items were represented solely by learnable embeddings; (2) DMM-NoSeq, where the LSTM-Attention module was replaced by a static

average of user history; and (3) DMM-NoAttn, where the attention mechanism was omitted, utilizing the final hidden state of the LSTM directly.

The results of the ablation study indicate that the Audio module (CNN) contributes most significantly to the diversity and novelty of recommendations, while the Sequential module (LSTM) is critical for accuracy. Removing the audio component (DMM-NoAudio) resulted in a 8.5% drop in NDCG@20, confirming that acoustic features provide essential information that is orthogonal to interaction data. The impact of sequential modeling was particularly pronounced; DMM-NoSeq suffered a 14.2% drop in performance. This highlights that music preference is strongly context-dependent: knowing a user's musical preferences alone is insufficient without also understanding the contextual conditions under which those preferences are expressed. The attention mechanism also proved vital; DMM-NoAttn performed worse than the full model, suggesting that dynamically weighing historical interactions (e.g., focusing on the last 5 songs more than songs from a year ago) is crucial for capturing the user's current mood.

Finally, we analyzed the model's sensitivity to key hyperparameters, specifically the dimensionality of the latent space d and the length of the input sequence L. We varied d in the range $\{32,64,128,256\}$. Performance improved consistently as d increased from 32 to 128, enabling the model to capture more nuanced features. However, increasing d to 256 yielded diminishing returns and significantly increased training time, leading us to select $d = 128$ as the optimal trade-off. Similarly, sequence length L was tested from 10 to 100. We found that $L = 50$ was sufficient to capture relevant history; utilizing longer sequences introduced noise and diluted the attention weights, slightly degrading the NDCG scores.

5. Discussion

The empirical results presented in the preceding section demonstrate that the DMM-HyRec system outperforms existing state-of-the-art methods. In this section, we provide a deeper theoretical interpretation of these findings by examining the underlying synergistic mechanisms and their broader implications for digital music recommendation. The effectiveness of DMM-HyRec does not stem merely from the use of deep learning techniques, but rather from an architectural design that aligns with the multifaceted nature of music preference. Users engage with music for diverse reasons: some are driven by intrinsic properties of the tracks, others by situational context, and still others by evolving preferences over time.

5.1 Analyzing the Synergy of Acoustic and Temporal Features

The significant performance gap between DMM-HyRec and unimodal baselines (CNN-Audio and BPR-MF) verifies our primary hypothesis: acoustic content features and collaborative behavioral data carry orthogonal yet complementary information. Purely collaborative models, such as BPR-MF and NeuMF, treat music tracks as abstract identifiers and rely entirely on the "wisdom of the crowd." Consequently, these models are highly vulnerable to data sparsity and the cold-start problem. Conversely, purely content-based models like CNN-Audio capture the "what" (e.g., tempo, instrumentation, genre) but lack the context of "who" and "when." For instance, a user may appreciate fast electronic music generally but restrict their listening to workout sessions—a contextual nuance that a content-only model fails to capture.

DMM-HyRec bridges this gap via a gated multi-modal fusion mechanism. The model captures complex nonlinear relationships between content and context by projecting high-level acoustic embeddings (from ResNet) and dynamic user state vectors (from Bi-LSTM) into a shared latent semantic space. For example, the model can learn that for User A, acoustic similarity drives sequential behavior (maintaining genre consistency), whereas for User B, novelty is prioritized, leading to transitions between acoustically distinct styles. Furthermore, the attention mechanism enhances this capability by enabling the model to focus on semantically relevant portions of a user's history. If the candidate song is a jazz track, the attention weights will automatically prioritize previous jazz tracks in

the sequence—regardless of their temporal distance—thereby providing a more accurate estimation of user preference than simple recency-based heuristics.

5.2 Mitigation of the Cold-Start Problem

A significant contribution of this work is the effective mitigation of the item cold-start problem, which remains a longstanding bottleneck in industrial recommender systems. Traditional systems typically require a 'warm-up' period, necessitating accumulated interactions to generate reliable recommendations. In our ablation study, excluding the audio component resulted in a substantial performance decline for less popular items. By leveraging raw audio spectrograms, DMM-HyRec generates dense latent representations for new tracks immediately upon upload, independent of historical user interaction.

The CNN functions as a universal feature extractor that maps acoustic signals to preference-related representations. When a new track is introduced into the system, the CNN analyzes its spectrogram and positions it in proximity to acoustically similar tracks within the latent space (e.g., similar timbre, rhythm, or harmonic structure). Consequently, users who have historically exhibited preferences for that region of the latent space can be recommended the new track immediately. This capability carries important economic implications for digital music platforms, as it facilitates the discovery of emerging artists and promotes a more balanced exposure distribution. Rather than reinforcing a “rich-get-richer” dynamic in which already popular artists dominate user attention, the system supports a recommendation environment that more strongly reflects intrinsic content characteristics.

5.3 Industrial Relevance and Scalability Considerations

When academic research focuses on marginal improvements in predictive accuracy, real-world deployment involves trade-offs between effectiveness, computational efficiency, and scalability. The DMM-HyRec architecture is designed with industrial constraints in consideration. Although training the coupled CNN–LSTM network requires substantial computational resources and a high-performance GPU environment, the inference phase is computationally efficient. The acoustic embeddings for the entire music catalog can be precomputed offline and subsequently stored in a high-performance vector database.

In the recommendation serving stage, generating a top-K list does not require recomputing the spectrograms. Instead, the system retrieves the precomputed user vector and performs an efficient ANN search within the item vector space. Technologies such as HNSW graphs or FAISS enable this search across millions of items within milliseconds. Such offline feature extraction, combined with efficient online retrieval, ensures that DMM-HyRec performs effectively on catalogs containing tens of millions of tracks, thereby providing the low-latency responses required by modern streaming systems.

6. Conclusion

The main contribution of the study is threefold. First, we constructed an end-to-end differentiable architecture that integrates the acoustic feature learning of raw audio spectrograms through residual CNNs with sequential user behavior modeling through attention-based Bi-LSTMs. The combination of these two aspects can capture both the intrinsic properties of music as well as the extrinsic context in which it is consumed. Second, DMM-HyRec was extensively evaluated through an empirical study on a large composite dataset which demonstrated that, compared with state-of-the-art baselines, including Neural CF and Self-Attentive Sequential approaches, the model achieved more than 15% improvement in NDCG@20. Third, both theoretical analysis and empirical evidence indicate that the model effectively alleviates the item cold-start problem and provide accurate and relevant recommendations for newly appearing content with no more than an audio analysis.

Despite these advances, several directions remain for future research. In the current framework, semantic representation is primarily derived from acoustic signals and interaction behavior. Future work

could incorporate additional textual modalities, such as lyrics and user-generated reviews, to further enrich semantic modeling. Leveraging pre-trained language models for textual understanding may enable the system to capture higher-level thematic and sentiment-related concepts, thereby enhancing personalization.

Additionally, although LSTMs are effective for sequential modeling, recent Transformer-based architectures have demonstrated superior performance in capturing long-range dependencies in NLP. For example, replacing LSTM modules with bidirectional Transformer encoders, such as those used in BERT4Rec, could be explored to further improve the modeling of complex and long-term user preference dynamics. Finally, extending the framework to cross-domain recommendation scenarios, such as suggesting concert tickets or merchandise based on a user's music listening history, represents a promising direction for enhancing ecosystem integration and generating additional economic value within digital entertainment platforms.

References

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., & Lamere, P. (2011). The million song dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference* (pp. 591-596).

Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., & Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4), 668-696.

Celma, O. (2010). Music recommendation and discovery in the long tail. *Springer Science & Business Media*.

Chen, T., Xu, W., Bates, S., & Recht, B. (2023). A statistical perspective on retrieval-augmented generation. *arXiv preprint arXiv:2309.15269*.

Cheng, H. T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., … & Dean, J. (2016). Wide & deep learning for recommender systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems* (pp. 7-10).

Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). Convolutional recurrent neural networks for music classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2392-2396).

Cremonesi, P., Koren, Y., & Turrin, R. (2010). Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the Fourth ACM Conference on Recommender Systems* (pp. 39-46).

Deldjoo, Y., Schedl, M., Cremer, P., & Knees, P. (2018). Content-based multimedia recommendation systems: Definition and application domains. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization* (pp. 289-290).

Dieleman, S., & Schrauwen, B. (2014). End-to-end learning for music audio. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6964-6968).

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12), 61-70.

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6), 602-610.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770-778).

He, X., & Chua, T. S. (2017). Neural factorization machines for sparse predictive analytics. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 355-364).

He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T. S. (2017). Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 173-182).

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., … & Slaney, M. (2017). CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 131-135).

Hidasi, B., Karatzoglou, A., Baltrunas, L., & Tikk, D. (2015). Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.

Jacobson, K., Murali, V., Newett, E., Whitman, B., & Yoner, R. (2020). Music personalization at spotify. In *Proceedings of the 14th ACM Conference on Recommender Systems* (pp. 649-649).

Kang, W. C., & McAuley, J. (2018). Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)* (pp. 197-206).

Koren, Y. (2008). Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 426-434).

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8), 30-37.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

Li, X., Wang, M., & Sun, T. P. (2023). A survey on deep learning for sequential recommendation. *Journal of Software*, 34(1), 77-103.

Liu, Z. (2026). Supply chain management in content distribution for Malaysian Chinese media services in Southeast Asia. Journal of Logistics, Informatics and Service Science, 13(2), 230 – 245. https://doi.org/10.33168/JLISS.2026.0213.

Lippens, S., Martens, J. P., Leman, M., Baets, B. D., Meyer, H., & Tzanetakis, G. (2004). A comparison of human and automatic classification of semantic music descriptors. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics* (pp. 5925-5930).

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. *In Proceedings of the 14th python in science conference* (pp. 18-25).

Nguyen, H. S. (2026). Sustainable social media advertising as an informatics-enabled service system: Evidence from Facebook users in Vietnam. *Journal of Logistics, Informatics and Service Science*, *13*(2), 212–229. https://doi.org/10.33168/JLISS.2026.0212.

Oramas, S., Nieto, O., Barbieri, F., & Serra, X. (2017). Multi-label music genre classification from audio, text, and images using deep hybrid neural networks. *arXiv preprint arXiv:1707.04916*.

Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2009). BPR: Bayesian personalized

ranking from implicit feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence* (pp. 452-461).

Schedl, M., Zamani, H., Chen, C. W., Deldjoo, Y., & Elahi, M. (2018). Current challenges and visions in music recommender systems. *International Journal of Multimedia Information Retrieval*, 7(2), 95-116.

Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., & Jiang, P. (2019). BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (pp. 1441-1450).

Van den Oord, A., Dieleman, S., & Schrauwen, B. (2013). Deep content-based music recommendation. In *Advances in Neural Information Processing Systems* (pp. 2643-2651).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998-6008).

Wang, X., He, X., Wang, M., Feng, F., & Chua, T. S. (2019). Neural graph collaborative filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 165-174).

Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, 52(1), 1-38.