Evolution and Fairness Challenges of Content-Based and Multimodal Music Recommendation Systems

Yaoyu Zhang

{yaoyuzhang22@gmail.com}

School of Journalism and Communication, Tianjin Normal University, Tianjin, Tianjin, 300387, China

Abstract. The contradiction between massive music libraries and users' limited attention spans has become prominent, making recommendation systems a core bridge connecting content and consumers. The technological advancement of music recommendation systems, which is evolving from traditional collaborative filtering to multimodal deep learning, is the main topic of this study. It also examines the difficulties in algorithmic fairness that result from these technological advancements. This study divides and contrasts previous research using fundamental technologies and technical stages. Research indicates that technology has made progress in using state-space models (SSMs) for continuous-time modeling; the improvement in technical accuracy is accompanied by increased popularity bias, gender imbalance, and two-sided market unfairness, and existing offline evaluation metrics are insufficient to reflect long-term user satisfaction. This paper provides a theoretical basis and reference for developing a next-generation music recommendation system that balances accuracy, diversity, and social responsibility.

Keywords: Music Recommendation System; Multi-modal Deep Learning; Algorithmic Fairness; Popularity Bias; State Space Models

1 Introduction

With the popularization of mobile Internet, music streaming platforms such as Spotify and Deezer have completely changed the global music consumption pattern. Faced with hundreds of millions of music libraries, users face a serious problem of "selection overload", which makes Recommender Systems (RS) no longer just an auxiliary tool, but a key "gatekeeper" that determines content distribution traffic and user retention [1]. Mokoena and Obagbuwa's review further confirms that Artificial Intelligence (AI) automation technology plays a decisive role in improving consumers' willingness to subscribe and platform loyalty, and is directly related to the commercial vitality of the platform [2]. However, with the commercial success of recommendation algorithms, their potential negative effects have gradually emerged. How to ensure the fairness and ecological health of algorithms while pursuing commercial efficiency has become the focus of attention in academia and industry.

Looking back at the development of music recommendation systems, early research mainly

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365225

relied on Collaborative Filtering (CF) and tag-based metadata matching. Although these methods solved the recommendation problem to some extent, they faced a serious cold start dilemma. In recent years, with the explosion of deep learning (DL) technology, the use of convolutional neural networks (CNN) and recurrent neural networks (RNN) to process raw audio signals and sequence data has become mainstream [3]. In particular, in 2024-2025, the introduction of state space model (SSM) and multimodal fusion technology has further pushed recommendation systems to a new stage of intelligence that can perceive context and understand semantics. Although the technology is changing rapidly, the existing review studies mostly focus on improving the accuracy of algorithms and lack a systematic review of the proposition of "how technological evolution affects the fairness of algorithms", especially the discussion on fairness in cutting-edge fields such as multimodal and continuous time modeling is still insufficient.

This paper aims to fill this gap, focusing on how music recommendation systems have evolved from "looking at tags" to "comprehensive understanding," and analyzing the popularity bias, gender imbalance, and bilateral market trade-offs that arise during this intelligent process. The paper is structured as follows: Section 2 details the technological evolution path and core classifications of music recommendation systems; Section 3 compares the practical differences among mainstream platforms; Section 4 deeply analyzes existing fairness challenges and research gaps; and Section 5 summarizes the entire paper and looks to future directions.

2 Technological Evolution of Music Recommendation Systems

The technological evolution of music recommendation systems can be clearly divided into three stages, showing an evolutionary logic from discrete to continuous and from single-modality to multi-modality.

The first stage was the early exploration based on content and collaborative filtering (2013-2019). Early recommendation systems mainly relied on collaborative filtering, but were plagued by the cold start problem. van den Oord et al. pioneered the use of CNNs to process Mel-spectrograms, directly extracting latent factors from audio signals, effectively alleviating the problem of new song recommendation [4]. Schedl's review further established the core position of deep learning in extracting audio content features [5].

The second stage was the maturation of serialization and hybrid modeling (2020-2023). With the accumulation of user interaction data, the research focus shifted to sequence modeling. Anand et al. combined CNNs with Long Short-Term Memory (LSTM) networks to not only extract static audio features, but also capture the temporal dynamics within the audio signal (such as the evolution of melody and rhythm), thereby improving the accuracy of content-based recommendations [6]. Meanwhile, hybrid recommendation models have emerged. The model proposed by Bahadure et al. combines popularity, personalized collaborative filtering and content-based filtering, and uses K-Means clustering to optimize the recommendation results, significantly improving the robustness of the system [3].

The third stage is context-aware, multimodal and continuous time modeling (2024-2025). The current cutting-edge research is dedicated to solving more complex dynamic interaction problems. Mateos and Bellogín proposed a new perspective of using context as an "interaction prerequisite" [7]. Li et al. designed the SARE module to realize personalized context fusion [8]. In terms of modeling technology, Xiao et al. introduced the state space model (SSM/Mamba) to construct SS4Rec, which solved the defect of traditional RNN/Transformer in not being able to handle irregular time intervals and realized continuous time series modeling [9]. In addition, the vMF-exp algorithm proposed by Benda et al. uses the von Mises-Fischer distribution to efficiently sample in a library of tens of millions of songs, solving the exploration bottleneck of large-scale action space [10].

3 Classification and Analysis of Core Technologies

3.1 Classification of Current Core Technologies for Music Recommendation

The core technologies of current music recommendation are mainly divided into five categories: deep content perception, sequence and time modeling, multimodal and multi-criteria decision making, context perception, and fairness perception. First, deep content perception technology is committed to directly analyzing audio waveforms and spectra through CNN or white-box parameter search to completely solve the cold start problem. For example, the white-box audio parameter search mechanism proposed by Yang et al. enhances the interpretability of the model [11], while the BiMMuDa dataset constructed by Hamilton et al. provides an important benchmark for symbolic melody analysis [12]. Although such methods have high computational costs and are difficult to capture social associations. Second, for the dynamic evolution of user interests, sequence and time modeling technology uses architectures such as RNN, Transformer and SSM to capture subtle changes. For example, the SS4Rec model effectively handles the decay and transfer of interests through continuous time modeling [9]. Although there are still challenges in long-term modeling, it greatly improves the accuracy of short-term prediction. On this basis, multimodal and multicriteria decision-making technologies have further expanded the data dimensions, integrating physiological signals such as audio, lyrics, and even facial expressions and speech (as in the research of Faye et al. [13]), and combining them with MCDM frameworks (such as Liu's work [14]) to achieve highly personalized recommendations. However, the alignment of multi-source heterogeneous data is still a technical challenge. Meanwhile, context-aware technologies take into account time, location, and activity status. For example, the SARE module designed by Li et al. can dynamically adapt to user scenario preferences, which significantly improves the experience while also bringing the risk of privacy leakage [8]. Finally, with the increasing attention to ecological health, fairness-aware technologies have become an important direction, balancing accuracy and fairness by introducing reordering or counterfactual learning mechanisms. Representative studies include the FULTR algorithm optimized end-to-end by policy gradients by Yadav et al. [15], and the empirical study by Ferraro et al. on the decisive role of reordering algorithms in alleviating gender imbalance in the long term [16].

3.2 Comparison of Mainstream Music Recommendation Systems in Practice

Mainstream music streaming platforms each have their own focus in terms of technology selection. The following comparison is based on case studies from relevant literature:

Table 1 summarizes and compares the recommendation system technology paths, user experience design, and fairness optimization strategies of the three major streaming media platforms: Spotify, Deezer, and Netflix. Different platforms have developed differentiated practices based on their own business scenarios. Spotify focuses on exploring content diversity through hybrid models and reinforcement learning; Deezer relies on collaborative filtering and vector search to ensure a smooth listening experience; and Netflix uses deep learning and causal inference to improve the accuracy and long-term value of video recommendations.

Table 1. Comparison of Recommendation System Technologies and Strategies among Spotify, Deezer, and Netflix

Dimension	Spotify	Deezer	Netflix（Video benchmark）
Core technologies	Hybrid Model + Reinforcement Learning (RL): Applying collaborative filtering (CF) and audio analysis; leveraging reinforcement learning (RL) to balance exploration and exploitation.	Collaborative Filtering + Large-Scale Vector Search: Applying SVD embedding; using the vMF-exp algorithm for efficient sampling in millions of directories.	Deep Learning + Causal Inference: Apply autoencoders/RNNs; use counterfactual evaluation to correct positional biases.
User experience	Discovery and Diversity Guide users to long-tail content through "Weekly Discovery"; balance active exploration with passive recommendations.	Instant gratification and the flow experience: Focus on a seamless “Flow” mode; vMF-exp ensures millisecond-level latency and a balance between correlation and randomness.	Explainability and Visualization: Provide explanations such as "Because you've viewed..."; use personalized artwork to increase click-through rates.
Application scenarios	Daily Mix, Release Radar, Situational Playlist	Real-time flow, "Mixes inspired by..." playlist.	Homepage sorting, video cold start recommendations.
Fairness strategy	Prevent information cocoons. Monitor consumption diversity; discover the correlation between diversity and high retention rates, and use this to guide algorithm optimization.	Addressing bias in the Large Action Space: Resolves sampling bottlenecks in the Large Action Space to prevent ignoring long-tail content.	Correcting Metric Misalignment: Shifting the focus from offline accuracy to long-term user value to fix the inconsistency between offline and online metrics.

4 Existing challenges and research gaps

First, there is the popularity bias and the Matthew effect. The Matthew effect is the phenomenon that popular content gets more exposure and long-tail content is further suppressed. The algorithm naturally tends to recommend popular content with high playback volume, which directly leads to a serious lack of exposure opportunities for long-tail musicians. Ungruh et al. pointed out that although alleviating this bias may slightly reduce the recommendation accuracy, it can significantly improve users' long-term satisfaction with discovering new songs[17]. Second, there is the gender imbalance and feedback loop. The feedback loop refers to the vicious cycle in which the algorithm generates recommendations based on users' historical clicks, thereby reinforcing the initial bias. Ferraro et al.'s research shows that if this loop is not intervened, the exposure rate of female artists will continue to deteriorate over time and is difficult to improve[18]. There is also a clear conflict of interest in the two-sided market. The two-sided market means that the platform serves both users and artists. The interests of the two groups are naturally different. Mehrotra et al. pointed out that simply optimizing the content relevance on the user side can easily damage the fairness of artists. How to find a Pareto optimal solution that takes both into account is still a problem that the industry urgently needs to solve[19]. Since the large language models (LLM) have been applied so widely, new risks of bias have naturally arisen, and Sah et al. have clearly and convincingly warned that stereotypes present in training data can be amplified by LLM, leading to discriminatory recommendation results, while also pointing out that current evaluation frameworks lack effective countermeasures [20]. Moreover, there is a well-documented problem of misalignment between offline and online evaluation: Steck et al.'s work at Netflix shows that models performing well on offline tests do not necessarily translate into good online user satisfaction in real-world settings. Hence, a multi-dimensional evaluation system that considers long-term user value is urgently needed.[21].

Future work on music recommendation is expected to proceed along three clear and natural lines: first, end-to-end fairness learning with policy gradient optimization for fairness objectives, second, improving the interpretability of recommendations via large models to thereby increase users' acceptance of diverse content, and third, designing a dynamic fairness mechanism using continuous-time models.

5 Conclusion

This study maps a clear evolutionary trajectory of music recommendation systems: from early single-framework collaborative filtering to modern multimodal, continuous-time modeling architectures. Today's core technologies enable deep, nuanced understanding of audio content, alongside dynamic, real-time perception of user context. Yet these gains in technical accuracy have not automatically translated to fairer system performance. On the contrary, popularity bias, gender representation imbalance, and conflicting interests across two-sided markets are all being subtly amplified within the algorithm's "black box".

The review serves first as a clear guide for technology selection from CNNs to SSMs, but far more importantly, it makes a compelling argument for introducing fairness restrictions at the earliest stage of algorithm design, and then naturally follows up with detailed discussions of comparative re-ranking and counterfactual learning as effective mitigation techniques.

References

[1] Anderson, A., Maystre, L., Anderson, I., Mehrotra, R., Lalmas, M.: Algorithmic effects on the diversity of consumption on Spotify. In: Proceedings of the Web Conference 2020 (WWW). pp. 2155–2165 (2020)

[2] Mokoena, N., Obagbuwa, I.C.: An analysis of artificial intelligence automation in digital music streaming platforms for improving consumer subscription responses: A review. Frontiers in Artificial Intelligence. 7, Article 1515716 (2025)

[3] Bahadure, N.B., Vishwakarma, M., Kurhekar, M.P., Patni, J.C., Patil, P.D.: Machine learning-based music classification and recommendation system from Spotify. International Journal of Computer Information Systems and Industrial Management Applications. 17, pp. 143–154 (2025)

[4] van den Oord, A., Dieleman, S., Schrauwen, B.: Deep content-based music recommendation. In: Advances in Neural Information Processing Systems. 26, pp. 2643–2651 (2013)

[5] Schedl, M.: Deep learning in music recommendation systems. Frontiers in Applied Mathematics and Statistics. 5, Article 44 (2019)

[6] Anand, R., Sabeenian, R.S., Gurang, D., Rubeena, S.: AI based music recommendation system using deep learning algorithms. IOP Conference Series: Earth and Environmental Science. 785, 012013 (2021)

[7] Mateos, P., Bellog ń, A.: A systematic literature review of recent advances on context-aware recommender systems. Artificial Intelligence Review. 58, Article 20 (2024)

[8] Li, J., Sun, P., Jiang, C., Ma, W., Ai, Q., Zhang, M.: A situation-aware enhancer for personalized recommendation. In: Proceedings of the 29th International Conference on Database Systems for Advanced Applications (DASFAA). pp. 1–15 (2024)

[9] Xiao, W., Wang, H., Zhou, Q., Wang, Q.: SS4Rec: Continuous-time sequential recommendation with state space models. arXiv preprint arXiv:2502.08132 (Accepted to IEEE TKDE) (2025)

[10] Bendada, W., Salha-Galvan, G., Hennequin, R., Bontempelli, T., Bouab ça, T., Cazenave, T.: Exploring large action sets with hyperspherical embeddings using von Mises-Fisher sampling. In: Proceedings of the 42nd International Conference on Machine Learning (ICML). pp. 1–12 (2025)

[11] Yang, Y., Jin, Z., Barnes, C., Finkelstein, A.: White box search over audio synthesizer parameters. In: Proceedings of the 24th International Society for Music Information Retrieval Conference (ISMIR). pp. 190–197 (2023)

[12] Hamilton, M., Clemente, A., Hall, E., Pearce, M.: The Billboard Melodic Music Dataset (BiMMuDa). Transactions of the International Society for Music Information Retrieval. 7(1), pp. 113–128 (2024)

[13] Faye, P., Sonar, S., Shahare, S., Ambhore, R., Chafle, V.: Emotion based music recommendation system. International Journal on Advanced Electrical and Computer Engineering. 14(01), pp. 157–160 (2025)

[14] Liu, J.: Aggregating contextual information for multi-criteria online music recommendations. IEEE Access. 13, pp. 8790–8805 (2025)

[15] Yadav, H., Du, Z., Joachims, T.: Policy-gradient training of fair and unbiased ranking functions. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1044–1053 (2021)

[16] Ferraro, A., Ekstrand, M.D., Bauer, C.: It's not you, it's me: The impact of choice models and ranking strategies on gender imbalance in music recommendation. In: Proceedings of the 18th ACM Conference on Recommender Systems (RecSys). pp. 884–889 (2024)

[17] Ungruh, R., Dinnissen, K., Volk, A., Pera, M.S., Hauptmann, H.: Putting popularity bias mitigation to the test: A user-centric evaluation in music recommenders. In: Proceedings of the 18th ACM Conference on Recommender Systems (RecSys). pp. 169–178 (2024)

[18] Ferraro, A., Serra, X., Bauer, C.: Break the loop: Gender imbalance in music recommenders. In: Proceedings of the 2021 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR). pp. 249–254 (2021)

[19] Mehrotra, R., McInerney, J., Bouchard, H., Lalmas, M., Diaz, F.: Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness and satisfaction in recommendation systems. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM). pp. 2243–2251 (2018)

[20] Sah, C.K., Xiaoli, L., Islam, M.M.: Unveiling bias in fairness evaluations of large language models: A critical literature review of music and movie recommendation systems. International Journal of Innovative Science and Research Technology. 8(12), pp. 2084–2093 (2023)

[21] Steck, H., Baltrunas, L., Elahi, E., Liang, D., Raimond, Y., Basilico, J.: Deep learning for recommender systems: A Netflix case study. AI Magazine. 42, pp. 7–18 (2021)