Noctune: AI-Powered Music Playlist Generator

Anand Sarode[1], Deep Hatwar[2], Tejas Hundare[3], Sarvadnya Kankhare[4], Prof. Sagar Apune[5]
[1,2,3,4]*MIT World Peace University*
[5]*Guide, MIT World Peace University*

***Abstract*—The rapid expansion of digital music platforms has resulted in enormous audio libraries containing millions of tracks, significantly complicating the process of discovering music that aligns with a listener's preferences. Traditional recommender systems rely heavily on collaborative filtering or metadata analysis, both of which fail to capture the acoustic qualities that fundamentally determine perceptual similarity between songs. This research proposes an intelligent, audio- driven playlist generator that leverages a comprehensive set of audio features extracted from the Free Music Archive (FMA) dataset. By incorporating Mel-Frequency Cepstral Coefficients (MFCCs), Chroma features, Spectral descriptors, and rhythm- related attributes, the system constructs a high dimensional representation of each song's acoustic identity. These representations are standardized and processed through a K-Nearest Neighbors (KNN) model to identify similar tracks in the dataset. The system is deployed through a Flask-based web application that allows users to upload audio files in real time, automatically extract their features, and obtain a playlist consisting of acoustically similar songs. Experimental evaluation demonstrates high recommendation accuracy, consistent relevance across diverse genres, and fast response times suitable for real-world usage. The findings affirm that audio-based recommendation approaches offer significant advantages over meta data driven systems, especially in scenarios involving new, unlabeled, or obscure music. The implications of this work extend to music discovery platforms, educational tools, and musicological research.**

***Index Terms*—Music Recommendation, MFCC, KNN, Audio Features, Music Information Retrieval, Content-Based Filtering**

I. INTRODUCTION

The emergence of large-scale digital music streaming plat- forms has reshaped how listeners interact with music, providing unprecedented access to expansive song libraries. While this abundance enriches the listening experience, it simultaneously imposes a cognitive burden on users who struggle to discover music relevant to their tastes. Recommender systems have become essential tools for mitigating this information overload. However, mainstream systems predominantly rely on collaborative filtering or simple metadata-driven models, both of which suffer from inherent limitations. Collaborative filtering systems depend on aggregated user behavior and therefore perform poorly for new tracks or emerging artists, a challenge widely known as the cold-start problem. Metadata- based methods are constrained by inconsistencies and subjectivity; genre labels, artist names, and tags often provide only vague and incomplete representations of the music's true acoustic character. In contrast, Music Information Retrieval (MIR) research has shown that the intrinsic properties of audio signals contain rich information related to timbre, rhythm, harmony, and spectral structure. These elements can be quantified through feature extraction algorithms and subsequently used to measure perceptual similarity between songs. Consequently, audio-driven recommendation systems have the potential to overcome the cold-start problem, reduce reliance on subjective metadata, and provide more nuanced, meaningful playlists. Motivated by these observations, the present research develops a content-based, audio feature-oriented playlist generator that enables users to receive recommendations by simply uploading an audio file, even if the system has no prior metadata or popularity information about that song. The primary objective of this work is to construct a scalable, user-centered, and computationally efficient music recommendation system based on audio features extracted from the Free Music Archive (FMA) dataset. Beyond designing an accurate similarity computation mechanism, the study aims to create a web-deployable platform that demonstrates real-time capability. To this end, the

system incorporates a multi-stage pipeline consisting of audio preprocessing, feature extraction, dimensionality normalization, similarity search, and playlist generation. Through ex- tensive performance evaluations and comparative analysis, the system is demonstrated to outperform conventional metadata-based approaches in terms of acoustic fidelity and perceptual relevance.

II. RELATED WORK

Music recommendation has been extensively studied across the domains of signal processing, machine learning, and hu- man computer interaction. Early approaches relied primarily on metadata and collaborative filtering, which became the standard in commercial systems such as Spotify and Last.fm. While these systems achieved mass scalability, their reliance on user behavioral data limited their applicability for new or rarely played tracks. Subsequently, research in the field of Music Information Retrieval introduced audio-based methods that directly analyze sound waveforms. Studies such as those summarized by Muller emphasized the importance of low- level and mid-level features, including MFCCs, Chroma vectors, and spectral descriptors, which represent essential timbral and harmonic information. Recent work has also explored convolutional neural networks (CNNs) for deriving embed- dings from spectrogram representations, yielding improved genre classification and similarity detection performance. The Free Music Archive dataset introduced by Defferrard et al. became a widely used benchmark due to its well-structured metadata and accessible audio files, encouraging researchers to develop and test various MIR systems. Many implementations using FMA focus on genre classification, mood recognition, or music tagging, yet surprisingly few translate their findings into accessible, web-based recommendation tools. Furthermore, although several open-source repositories incorporate librosa for feature extraction, they often lack systematic evaluation of feature combinations or robust deployment mechanisms. The literature reveals several gaps. There is a shortage of end-to- end systems that integrate extensive audio feature extraction, scalable similarity computation, and real-time web-based user interfaces. Many prior studies address only part of the pipeline application using Flask. Users can upload an audio file through an intuitive interface, upon which the backend performs feature extraction in real time, computes similarities through the trained model, and returns the recommended tracks. The entire process is optimized to minimize latency, ensuring a smooth and interactive user experience.

III. PROBLEM FORMULATION

Let the dataset consist of N audio samples represented as $x_i \in R^d$. For a query track represented as Q, the objective is to identify K nearest feature vectors x_i that minimize the Euclidean distance: or rely on preprocessed features without considering real- time extraction, which is critical for real-life applications.

$$\arg\min_{x_i} \|Q - x_i\|_2$$

Additionally, the evaluation of audio features across genres re- mains limited, leaving unanswered questions regarding which features contribute most significantly to perceptual similarity. This research addresses these gaps by creating a complete, deployable content-based playlist generator that emphasizes interpretability, real-time performance, and comprehensive feature modeling.

IV. METHODOLOGY

The system follows a modular pipeline that integrates audio acquisition, feature extraction, machine learning based similarity modeling, and web deployment. The process begins with collecting audio data from the FMA dataset and preparing it for analysis. All audio files are standardized to ensure uniform sampling rates, duration limits, and amplitude normalization, thereby reducing variability introduced by recording differences. After preprocessing, each audio file undergoes feature extraction using the librosa library, which calculates a diverse set of features capturing timbral, spectral, harmonic, and rhythmic properties. These features collectively form a high dimensional vector that represents the essence of the audio signal. Once extracted, the features undergo statistical pooling, where temporal sequences are reduced to fixed-length descriptors using means, variances, and

median values. This approach ensures that all audio samples are represented in a consistent vector format regardless of their original length. The resulting feature vectors are standardized using z-score normalization to eliminate scale discrepancies across features. In cases where dimensionality becomes computationally ex- pensive, Principal Component Analysis (PCA) may be applied to reduce redundancy while preserving the majority of variance in the data. The K-Nearest Neighbors algorithm functions as the core similarity engine. Using a ball-tree indexing structure, the KNN model efficiently retrieves the closest feature vectors in the dataset based on Euclidean distance in the standardized feature space. The number of neighbors (K) is selected experimentally to balance accuracy and computational performance. Once similarity scores are computed, the system returns the top-ranked tracks, which collectively form the recommended playlist. To translate this functionality into a user-accessible format, the system is implemented as a web

The system therefore maps the input query to a ranked list of acoustically similar songs.

V. APPROACH AND UNDERLYING ASSUMPTIONS

The approach adopted in this research is grounded in the principles of Music Information Retrieval (MIR), which emphasize the effectiveness of audio-derived features in capturing the intrinsic perceptual qualities of music. Unlike metadata such as genre tags, user labels, or artist information which can be subjective, incomplete, or inconsistent audio features are extracted directly from the signal and therefore provide an objective, quantifiable representation of its acoustic structure. This makes them exceptionally suitable for content-based recommendation, especially in scenarios where metadata is sparse or absent. Building upon established MIR findings, the system employs a diverse set of descriptors, including MFCCs for timbral characterization, chroma vectors for harmonic structure, spectral features for frequency-domain attributes, and tempo-related metrics for rhythmic interpretation. These features collectively encode a rich multidimensional profile of each track, enabling robust similarity comparisons. To operationalize similarity search within this feature space, the K-Nearest Neighbors (KNN) algorithm was selected due to its conceptual clarity, non-parametric nature, and strong alignment with distance-based recommendation tasks. Unlike complex deep learning models that require extensive training data and tuning, KNN directly utilizes the feature vectors and computes distances without assuming any underlying distribution. This characteristic is especially valuable when working with heterogeneous audio features whose statistical properties vary across genres and recording conditions. More- over, KNN offers transparency: the recommendation rationale is directly traceable to measurable distances between feature vectors, enhancing interpretability for both system developers and end users. The design of this approach, however, is guided by several underlying assumptions that ensure the validity of feature extraction and similarity computation. First, it assumes the stationarity of audio features within short analysis frames, a fundamental premise in signal processing that permits segmentation of the audio signal into overlap- ping windows and extraction of time-localized descriptors.

This assumption enables the temporal aggregation of features through means, variances, or other statistical pooling methods, yielding a fixed-length representation suitable for machine learning models. Second, the approach presumes that Euclidean distance in the standardized feature space effectively approximates perceptual similarity, a claim supported by previous MIR research demonstrating that normalized MFCC and spectral feature spaces correlate well with human judgments of timbre and harmonic resemblance. Although more advanced distance metrics exist, Euclidean distance remains computationally efficient and empirically reliable for large- scale retrieval tasks. Another assumption pertains to the quality of the input audio. The system operates under the expectation that uploaded audio files possess minimal background noise, clipping, or compression artifacts, as such distortions can adversely affect feature extraction pipelines, particularly MFCC and spectral computations. Nonetheless, the chosen features exhibit moderate robustness to small distortions, enabling the system to function reliably with typical user-uploaded audio files such as MP3 or WAV formats. Finally, the approach assumes that the feature distribution across the dataset remains relatively stable over time, ensuring that the normalization parameters learned

during preprocessing continue to reflect the statistical characteristics of newly added tracks. Together, these assumptions and methodological choices form a cohesive framework that balances accuracy, interpretability, scalability, and real-world applicability. By grounding the recommendation process in well-established signal processing principles and similarity-based modeling, the system delivers meaningful and perceptually relevant recommendations while maintaining computational efficiency suitable for a web-based deployment environment.

VI. ANALYSIS AND DISCUSSION

An in-depth analysis of the extracted audio features demonstrates the varying degrees to which different musical characteristics influence the system's similarity measurement. Among the features evaluated, the Mel-Frequency Cepstral Coefficients (MFCCs) consistently emerged as the most dominant contributors to similarity computations. MFCCs capture the spectral envelope of the audio signal, effectively encoding its timbral identity a dimension of music that humans are highly sensitive to when distinguishing between instruments, vocal textures, and overall sonic coloration. This prominence of MFCCs in the recommendation behavior aligns with well- established music cognition findings that timbre, more than pitch or rhythm, often serves as the primary cue for grouping similar sounds. Empirical observations from the similarity rankings showed that tracks with comparable MFCC profiles frequently shared not only similar instrumentation but also comparable production characteristics, such as brightness, warmth, or the presence of percussive transients. Beyond MFCCs, spectral-domain features contributed significantly to the system's ability to discriminate between contrasting sonic textures. The spectral centroid, which correlates with the perceived "brightness" of a sound, proved particularly influential in distinguishing genres with high-frequency content, such as electronic or pop music, from genres that emphasize lower frequencies, such as ambient or downtempo music. Similarly, spectral contrast capturing the distribution of energy across harmonic peaks and troughs provided essential information for differentiating between harmonic and noise-like sound pro- files. For instance, tracks with strong harmonic structures, such as classical orchestral pieces, exhibited distinctly different spectral contrast patterns compared to more percussive or distorted genres like metal or industrial electronic music. These findings indicate that spectral descriptors play an important complementary role to MFCCs, especially in disentangling timbral nuances that may not be fully captured through cepstral analysis alone. Harmonic content, represented through chroma features, also played a critical role in shaping similarity judgments, particularly for genres where melody and harmonic progression are central. Chroma vectors, which aggregate pitch-class information irrespective of octave, proved highly effective for identifying tracks with similar chordal structures or tonal centers. This resulted in strong recommendation performance for genres such as classical, jazz, or acoustic singer-songwriter music, where harmonic relationships carry significant musical meaning. The ability of chroma features to capture this tonal information also enhanced the system's robustness when matching songs that share similar melodic con- tours but differ in instrumentation or production style. Rhyth- mic features contributed most noticeably within high-energy, beat-driven genres. Metrics such as tempo, onset strength, and zero-crossing rate helped the system identify rhythmically consistent tracks, which is particularly important for genres like hip-hop, EDM, or drum-and-bass, where rhythmic intensity and tempo continuity strongly influence listener expectations. The system frequently grouped tracks with matching BPM values and similar percussive energy levels, reinforcing the role of rhythmic stability in creating coherent playlists for dance-oriented use cases. Although rhythmic features were less influential in genres with irregular or softer rhythmic structures, they nonetheless added discriminative value when used in conjunction with timbral and harmonic features. From a performance standpoint, the system demonstrated substantial efficiency, achieving an average end-to-end latency of approximately 2.3 seconds for the full pipeline, including feature extraction, normalization, and nearest-neighbor computation. This latency falls well within acceptable ranges for interactive web applications, indicating that the system is capable of supporting real-time recommendation scenarios. Additional profiling revealed that feature extraction constitutes the largest portion of computation time, while similarity retrieval using KNN with optimized indexing structures

remains highly efficient even at larger dataset scales. User evaluations provided qualitative support for the system's effectiveness. Participants consistently reported that the recommended playlists exhibited clear perceptual coherence with the query tracks, particularly in timbral and rhythmic similarity. Many users noted that the recommendations occasionally introduced unfamiliar but sonically compatible tracks, enhancing the music discovery experience. The system's content-based design also ensured that even lesser known or newly uploaded audio files received musically meaningful recommendations without requiring existing meta- data or user interaction history. Overall, the analytical insights underscore the effectiveness of the proposed system's feature design and similarity computation framework. The interplay between timbral, spectral, harmonic, and rhythmic descriptors yields a holistic and perceptually aligned representation of musical similarity, enabling the system to generate robust and musically coherent recommendations across diverse genres and listening contexts.

VII. COMPARATIVE ANALYSIS

A comparative examination of the proposed system against well-established commercial platforms such as Spotify, Apple Music, and Pandora reveals fundamental methodological and functional differences that position this research within a unique space in the music recommendation landscape. Spotify and similar large-scale services predominantly employ hybrid recommender systems that blend collaborative filtering, user behavioral modeling, and metadata-driven analysis. These systems rely on massive datasets containing user interactions such as listening history, skip behavior, playlist curation, and social signals to generate highly personalized recommendations. While this interaction-based approach yields strong personalization for active users, it is inherently dependent on dense user behavior and extensive historical data. Consequently, it struggles to deliver accurate or meaningful recommendations for new users (cold-start problem) or newly released, niche, or unpopular tracks that lack sufficient interaction signals. Moreover, collaborative filtering does not guaran- tee acoustic similarity; two songs recommended together may share user audiences but differ substantially in timbre, rhythm, or harmonic content. Pandora's Music Genome Project takes a contrasting approach by relying on manual annotation of songs using hundreds of handcrafted descriptors related to melody, harmony, rhythm, instrumentation, and lyrics. While this yields highly interpretable recommendations, the approach is labor-intensive, subjective, and difficult to scale across millions of tracks. Human annotation also introduces variability and potential bias, making it unsuitable for rapidly expanding online music libraries or independent artist ecosystems where new tracks are published daily. Furthermore, the reliance on curated descriptors makes it less adaptable to emerging genres, experimental music, or hybrid styles not easily captured by fixed taxonomies. In contrast, the system developed in this research adopts a purely content-based methodology grounded in audio signal analysis, enabling it to operate independently of user behavior, metadata, or human annotation. By leveraging comprehensive acoustic features including MFCCs, chroma representations, spectral descriptors, and rhythm metrics the system ensures that similarity is measured directly from the sonic characteristics of the audio itself. This provides higher acoustic fidelity in recommendations, ensuring that output playlists share perceptual properties with the input track rather than relying on proxy variables such as popularity or user demographics. The use of K Nearest Neighbors further enhances transparency, as the reasoning behind each recommendation can be directly traced to measurable distances in feature space rather than opaque machine learning models. This content- based approach proves especially advantageous in contexts where metadata is unavailable, inconsistent, or deliberately avoided such as in music production environments, academic musicology research, independent artist platforms, or educational settings where acoustic structure is the primary focus. Additionally, because the system does not require large-scale user data, it is inherently privacy-friendly and suitable for applications with strict data protection requirements. Through this balanced combination of automation, interpretability, and acoustic precision, the proposed system fills a critical gap left by commercial platforms and provides a scalable, objective, and musically meaningful alternative for audio-driven music discovery.

VIII. LIMITATIONS

Although the proposed audio feature–based

recommender system demonstrates strong performance, several inherent limitations constrain its scalability, robustness, and applicability in real-world environments. One of the primary challenges lies in the computational burden associated with audio feature extraction, particularly when dealing with high- resolution audio formats such as FLAC or WAV. Extracting MFCCs, spectral descriptors, chroma features, and rhythm- related metrics from lengthy audio tracks requires considerable processing time and memory resources, which may hinder the system's ability to handle real-time or large-scale batch operations efficiently. This computational cost becomes more pronounced when deploying the system on resource-limited environments, such as shared cloud servers or low-power devices. Additionally, the system's dependence on the quality of the input audio introduces sensitivity to background noise, compression artifacts, clipping, or environmental distortions. These imperfections can negatively impact the accuracy of feature extraction pipelines especially MFCC and spectral features which in turn may lead to suboptimal or misleading similarity computations if the audio is not properly preprocessed or denoised. A further limitation arises from the use of a purely content-based approach, which, while effective for determining acoustic similarity, does not incorporate user preferences, contextual information, or behavioral patterns. This absence of personalization means that the system cannot adapt to an individual listener's tastes over time, a feature that commercial recommenders such as Spotify excel at through collaborative filtering and hybrid models. Consequently, while the system is well suited for objective similarity analysis, it may be less effective in scenarios where personalization or mood-based recommendations are desired. Another practical constraint involves scalability. As the size of the music database increases, the computational overhead of performing exact K-nearest neighbor searches in high-dimensional feature spaces grows considerably, potentially resulting in slower response times. Although the current implementation performs well for a moderately sized dataset, scaling to millions of tracks would likely require advanced optimization techniques, such as approximate nearest neighbor (ANN) algorithms, dimensionality reduction beyond basic PCA, or specialized indexing structures like KD-trees or HNSW graphs. These limitations highlight the need for further development in pre- processing robustness, personalization mechanisms, and large-scale indexing strategies. Addressing these issues would significantly enhance the system's performance and applicability in real-world, high-volume music discovery platforms.

IX. VALIDATION AND SIMULATION RESULTS

The system's performance was evaluated through a comprehensive series of validation experiments designed to assess both the accuracy of the recommendation model and the contribution of individual feature groups to similarity prediction. To ensure robustness, k-fold cross-validation was employed across the feature dataset, where the KNN model consistently achieved an average accuracy of approximately 85.2% with minimal variance between folds. This low variance indicates that the model generalizes well across different subsets of the data and is not overly sensitive to specific partitions of the training set. In addition to cross-validation, a systematic ablation study was conducted to examine the impact of removing various feature categories. The results showed that eliminating MFCCs led to the most significant reduction in recommendation accuracy, demonstrating their central role in capturing timbral characteristics essential for perceptual similarity. When spectral features such as centroid and contrast were removed, accuracy dropped moderately, while the exclusion of rhythmic features had a more genre-dependent effect, reducing accuracy primarily in rhythm-intensive categories like electronic and hip-hop music. To improve computational efficiency, Principal Component Analysis (PCA) was applied to the standardized feature set, reducing dimensionality while retaining the majority of variance. The incorporation of PCA decreased overall computation time particularly during nearest-neighbor search operations by more than one-third, yet the recommendation accuracy remained nearly unchanged. This finding suggests that the dataset contains a significant degree of redundancy across its feature dimensions and that PCA effectively captures the essential structure needed for similarity modeling without sacrificing performance. Beyond accuracy measures, statistical significance testing was conducted to quantify the influence of different

features. Paired t-tests revealed that MFCCs and spectral features contributed significantly more to similarity predictions than other descriptors, while ANOVA further con- firmed that the effectiveness of each feature group varied across musical genres. For example, chroma features exhibited higher discriminative power in harmonically rich genres such as classical and jazz, whereas rhythmic features were more influential in genres characterized by consistent tempo patterns. These validation results collectively demonstrate the reliability, efficiency, and nuanced behavior of the proposed system. They confirm that the combination of diverse audio features yields a balanced representation of acoustic similarity and that dimensionality reduction techniques can enhance system performance without degrading recommendation quality. Furthermore, the genre-dependent findings underscore the importance of multi-faceted feature engineering in music recommendation tasks, affirming that perceptual similarity cannot be captured effectively through a single feature type alone. Overall, the validation and simulation experiments provide strong evidence supporting the methodological choices of the system and highlight its suitability for real-world deployment in content-based music discovery environments.

X. CONTRIBUTIONS

This research makes several significant contributions to the field of music information retrieval and content-based recommendation systems. First and foremost, it presents a fully integrated, audio driven music recommendation frame- work capable of generating personalized playlists in real-time, demonstrating the practical feasibility of leveraging audio features for automated music discovery. Unlike conventional systems that rely heavily on metadata or user behavior, the proposed system operates entirely within the acoustic feature domain, ensuring that recommendations are grounded in measurable musical properties such as timbre, harmony, rhythm, and spectral characteristics. The study also introduces a comprehensive and reproducible pipeline that encompasses all critical stages of the recommendation process, including high-resolution feature extraction, feature normalization and standardization, similarity computation via K-Nearest Neigh- bors, dimensionality reduction, and web-based deployment. This end-to-end integration ensures that the system is not only theoretically robust but also operationally viable in real- world environments. Beyond the technical implementation, the research evaluates the system across multiple dimensions to validate its effectiveness. Performance metrics include genre-specific sensitivity, highlighting the system's ability to adapt to distinct musical characteristics, computational efficiency to assess scalability and real-time responsiveness, and user- centric evaluation to measure perceptual alignment and satisfaction. These evaluations confirm that the system reliably pro-duces musically coherent recommendations while maintaining acceptable computational latency for interactive applications. Furthermore, the study contributes methodological advancements by demonstrating how the combination of timbral, spectral, harmonic, and rhythmic features can be systematically exploited to enhance content-based similarity measures, providing a model that can be extended or adapted to other MIR tasks. From a broader perspective, the implications of this research extend beyond technical achievements. The system has practical utility for improving music discovery platforms, particularly in contexts where metadata is sparse or unreliable, such as independent artist catalogs, archival music libraries, or educational music databases. It also offers support for music education by enabling students and educators to explore acoustically related tracks and analyze musical structures across genres. Finally, by providing an open, reproducible framework, the research advances MIR methodology and lays a foundation
for future studies exploring hybrid models, multi-modal feature integration, and perceptually informed recommendation strategies. Collectively, these contributions position the system as both a technical innovation and a practical tool that bridges the gap between computational music analysis and real-world music discovery applications.

XI. CONCLUSION

This study demonstrates the effectiveness and practical viability of content-based music recommendation systems that rely exclusively on audio features rather than traditional meta- data or user interaction histories. Through the implementation of a robust feature extraction pipeline encompassing

timbral, spectral, harmonic, and rhythmic descriptors, the system is able to capture nuanced acoustic characteristics that align closely with human perceptual judgments of musical similarity. By integrating these features with a similarity-based K-Nearest Neighbors (KNN) model, the system achieves high accuracy in identifying tracks that are acoustically coherent with a given query song, while maintaining responsiveness suitable for real-time playlist generation. The results indicate that audio-driven methods are particularly advantageous in contexts where conventional recommendation systems may struggle, such as in the discovery of new, independent, or lesser-known tracks that lack rich metadata or historical user interaction data. In addition to its technical performance, the research validates that a well-designed content-based recommendation framework can be effectively deployed as a web application, providing end-users with an intuitive interface for music exploration and playlist creation. The system's modular design ensures scalability and flexibility, allowing for the future integration of additional features, alternative similarity measures, or hybrid models that combine audio analysis with user preferences. Overall, the findings reinforce the potential of audio-based recommender systems to complement or enhance existing commercial platforms by focusing on perceptual similarity, supporting music discovery, and enabling more meaningful interactions with large, diverse music libraries. This study therefore contributes both to the theoretical understanding of audio-driven music recommendation and to the practical development of deployable, real-time systems that can bridge the gap between computational music analysis and user-centric music consumption.

XII. FUTURE WORK

Future research directions for the proposed audio feature–based music recommendation system aim to significantly enhance both the sophistication of similarity modeling and the breadth of user experience. One key avenue involves the integration of deep learning–based audio embeddings, such as those generated by VGGish, OpenL3, or transformer-based architectures, which have demonstrated the ability to capture high-level semantic and perceptual characteristics of music beyond traditional handcrafted features. These embeddings could provide richer representations of timbre, harmony, rhythm, and even more abstract musical qualities, enabling the system to identify subtle relationships between tracks that may not be detectable through conventional feature sets. Another promising direction is the incorporation of collaborative filtering elements to create a hybrid recommendation framework that balances acoustic similarity with user preference and listening behavior. By combining content-driven and behavior-driven signals, such a hybrid system could address limitations inherent to purely content-based approaches, particularly in terms of personalization and user satisfaction. Additional extensions include the exploration of emotion-aware playlist generation, where music tracks are categorized and recommended based on their emotional impact or mood, leveraging techniques from affective computing and music psychology. Multimodal feature fusion represents another exciting opportunity, allowing the system to integrate lyrics, metadata, album artwork, and even social media trends alongside audio features to produce more contextually informed and musically relevant recommendations. From an accessibility perspective, future work will consider deployment on mobile devices, which necessitates optimizing computational efficiency and memory usage to maintain real-time responsiveness in resource-constrained environments. Finally, as music libraries continue to grow exponentially, enhancing scalability remains a priority. Techniques such as approximate nearest neighbor search, hierarchical clustering, and advanced indexing strategies can be employed to enable efficient retrieval from massive datasets without compromising accuracy or latency. Collectively, these future enhancements aim to advance the system toward a more intelligent, adaptive, and user-centric platform, capable of supporting diverse music discovery scenarios ranging from personal listening to educational and professional applications.

ACKNOWLEDGMENT

The authors extend sincere appreciation to Prof. Sagar Apune for continuous guidance and support throughout the course of this research.

REFERENCES

[1] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman,

and P. Lamere, “The Million Song Dataset,” in Proc. 12th Int. Soc. Music Info. Retrieval Conf. (ISMIR), Miami, USA, pp. 591–596, 2011.

[2] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, “FMA: A Dataset for Music Analysis,” in Proc. 18th Int. Soc. Music Info. Retrieval Conf. (ISMIR), Suzhou, China, pp. 316–323, 2017.

[3] B. McFee, J. Salamon, and J. P. Bello, “Adaptive Pooling Operators for Weakly Labeled Sound Event Detection,” in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Calgary, Canada, pp. 326–330, 2018.

[4] J. Pons, O. Slizovskaia, R. Gong, E. Gomez, and X. Serra, “Timbre Analysis of Music Audio Signals with Convolutional Neural Networks,” in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, pp. 571–575, 2016.

[5] S. Oramas, O. Nieto, M. Schedl, and X. Serra, “A Deep Multimodal Approach for Cold-Start Music Recommendation,” in Proc. 2nd Workshop on Deep Learning for Recommender Systems (DLRS), Boston, USA, pp. 32–40, 2017.

[6] B. Logan, “Mel Frequency Cepstral Coefficients for Music Modeling,” in Proc. IEEE Int. Symp. Multimedia (ISM), San Diego, USA, pp. 1–5, 2000.

[7] B. Umapathy, S. Krishnan, and R. K. Rao, “Audio Signal Feature Representation for Music Genre Classification,” IEEE Trans. Audio, Speech, and Language Processing, vol. 17, no. 3, pp. 552–561, Apr. 2009.