International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

A Deep Learning Framework for Emotion-Conditioned Personalized Music Recommendation

SAGAR BARGOTI , Dr. PRABHA NAIR

B. tech student, Department of IT, Noida Institute of Engineering Technology, Gr. Noida
Deputy HOD, Department of IT, Noida Institute of Engineering Technology, Gr. Noida

https://doi.org/10.55041/ijsmt.v2i5.178

Cite this Article: BARGOTI, S. (2026). A Deep Learning Framework for Emotion-Conditioned Personalized Music Recommendation. International Journal of Science, Strategic Management and Technology, *02*(05). https://doi.org/10.55041/ijsmt.v2i5.178

Abstract— While music acts as a powerful emotional regulator, traditional recommendation systems often fail to account for a user's immediate affective state, relying instead on static historical logs. We present **EmotionMuse**, a modular deep learning framework that bridges this gap by integrating real-time facial expression analysis with history-conditioned music suggestions. Our architecture utilizes a VGG-16 CNN, enhanced with Squeeze-and-Excitation attention, to achieve 87.2% accuracy in emotion classification on the FER-2013 dataset. These detections are mapped onto Russell's valence-arousal plane to generate 64-dimensional affective embeddings. These embeddings condition a Bidirectional LSTM (Bi-LSTM) model, which processes user listening sequences from the Million Song Dataset. Cross-dataset alignment is established through audio feature matching in Spotify's space to ensure theoretically grounded emotion-to-music correspondence. Experimental results demonstrate a Precision@10 of 0.791 and an NDCG@10 of 0.813, representing a performance gain of 5.8% over affect-blind baselines. Our system maintains an end-to-end latency of 94 ms, supporting real-time deployment on standard consumer hardware.

Keywords—Affective Computing, Bidirectional LSTM, Convolutional Neural Network, Emotion Recognition, FER-2013, Music Information Retrieval, Music Recommendation, Personalization, Valence-Arousal Space, Squeeze-and-Excitation Networks, Deep Learning, Facial Action Coding System, Attention Mechanisms.

I. INTRODUCTION

The proliferation of digital music streaming platforms has fundamentally transformed how individuals discover, consume, and interact with music across all demographics. Spotify alone reported over 600 million monthly active users in early 2024 [1], with Apple Music and YouTube Music recording similar growth trajectories, collectively serving catalogues that now exceed 100 million licensed tracks. At the core of each platform lies a recommendation engine that must select, from catalogues of tens of millions of tracks, a small set of songs that will genuinely resonate with the listener at a given moment in time. The economic implications are considerable: recommendation-driven engagement accounts for more than 30% of total streaming minutes on major platforms, making algorithmic personalisation a strategic priority for the industry [35].

Despite substantial algorithmic progress over the past decade, mainstream recommendation systems share a persistent structural limitation: they predominantly model user preference as a stationary or slowly evolving quantity derived from historical interaction logs. Collaborative filtering, matrix factorisation, and

their deep-learning successors all operate on the assumption that a user's future behaviour is best predicted by their past behaviour, aggregated without explicit reference to moment-to-moment context. In practice, however, a listener's musical taste is highly context-sensitive and emotionally driven. The same individual who seeks energetic, high-tempo tracks during a morning workout may gravitate toward contemplative ambient music during late-night study sessions, and toward socially vibrant pop during gatherings. Recommending based solely on historical behaviour—without awareness of the user's current affective state—constitutes a substantive failure of affective personalisation that reduces both user satisfaction and engagement [36].

Human emotion can be represented along two complementary theoretical axes. Discrete categorical models—most influentially Ekman's taxonomy [2]—identify anger, disgust, fear, happiness, sadness, surprise, and neutral as universally recognisable facial expressions with cross-cultural validity. Continuous dimensional models, specifically Russell's circumplex model of affect [3], represent emotional states as points within a two-dimensional plane defined by valence (the positive–negative hedonic dimension) and arousal (the high–low activation dimension). This dual-axis representation has proven particularly productive in Music Information Retrieval (MIR), where the emotional character of musical pieces has been systematically characterised through measurable audio features such as tempo, mode, key, energy, and spectral distributions [16]. The alignment between emotional states and musical attributes is well-established empirically: high-arousal positive emotions (excitement, joy) correlate strongly with fast tempo, major key, and high spectral energy, whereas low-arousal negative emotions (sadness, melancholy) associate with slow tempo, minor key, and softer dynamics.

Of the non-intrusive sensing modalities available for automated emotion recognition in consumer contexts—including facial expression analysis, speech prosody, electroencephalography (EEG), galvanic skin response (GSR), and accelerometer-based body movement—facial expression analysis is the most practically deployable at scale. It requires only the standard front-facing camera present in virtually every consumer device, imposes no additional hardware cost, is entirely non-contact, and provides rich affective signal at frame rates sufficient for real-time inference. Deep Convolutional Neural Networks (CNNs) have achieved strong performance on benchmark facial expression datasets [4], with recent models approaching and in some cases exceeding human-level agreement on well-curated test sets. Advances in efficient neural architecture design, quantisation, and hardware-accelerated inference have further reduced the computational barrier to deploying real-time CNN-based emotion recognition on commodity hardware, making camera-based affect sensing a practical component of consumer applications.

This paper presents EmotionMuse, a complete end-to-end system that infers the user's affective state from a brief webcam capture, encodes this state as a theoretically grounded learned emotional embedding in the valence-arousal space, and conditions a sequential Bi-LSTM music recommendation model on this embedding to generate a contextually appropriate ranked list of music suggestions. The system is designed with explicit attention to reproducibility, technical rigour, and honest evaluation. The primary scientific and engineering contributions are:

- A complete, modular four-stage pipeline—spanning facial image capture, CNN-based emotion classification with temporal smoothing, affective embedding generation, and BPR-trained Bi-LSTM sequential recommendation—implemented entirely with publicly available datasets and open-source libraries, and designed to be fully reproducible on consumer hardware at modest cost.
- A fine-tuned VGG-16 CNN augmented with Squeeze-and-Excitation (SE) channel-attention modules achieving 87.2% overall accuracy on the FER-2013 test set, accompanied by transparent per-class performance analysis, explicit treatment of class-imbalance mitigation strategies, and a principled comparison against published single-model benchmarks.
- A technically sound cross-dataset emotion-to-music alignment methodology using Spotify audio feature matching and cosine-similarity-gated k-nearest-neighbour label propagation from the DEAP stimulus corpus to the MSD catalogue, constituting a reproducible and theoretically grounded alternative to the unsupported direct cross-dataset label transfer employed in prior work.
- A temporal smoothing and hysteresis scheme for multi-frame emotion probability aggregation that

demonstrably reduces prediction instability during real-time operation, with quantified latency characteristics supporting deployment at approximately 10 Hz on mid-range GPU hardware.

- A comprehensive ablation study isolating the incremental contribution of each architectural component, and a candid critical discussion of dataset limitations, evaluation assumptions, demographic constraints, and practical deployment considerations relevant to applied affective computing.

The remainder of this paper is structured as follows: Section II reviews relevant prior work across traditional music recommendation, deep sequential recommendation, facial emotion recognition, and emotion-aware music systems, identifying the specific gaps addressed by this work. Section III describes the datasets employed and the cross-dataset coupling methodology. Section IV details the proposed system architecture across all four pipeline stages. Section V presents experimental results including emotion recognition accuracy, recommendation performance, and the ablation study. Section VI provides a critical discussion of findings, clarifications, limitations, reproducibility considerations, and practical implementation details. Section VII concludes the paper and identifies directions for future research.

II. Literature Review

A. Traditional Music Recommendation

Music recommendation research predates the deep learning era by more than two decades. Early systems relied on collaborative filtering (CF), which exploits statistical regularities in user-item interaction matrices to identify users with similar taste profiles and recommend items consumed by those peers. The seminal work of Koren et al. [5] demonstrated that Matrix Factorisation (MF) with Bayesian Personalised Ranking (BPR) provides strong baselines for implicit feedback settings by optimising a pairwise ranking objective directly. Simultaneously, content-based filtering (CBF) methods characterised music through audio-derived features—tempo, spectral centroid, zero-crossing rate, mel-frequency cepstral coefficients (MFCCs), and timbral texture descriptors—and recommended tracks with similar acoustic profiles to those previously enjoyed [6]. While CBF captures acoustic similarity effectively, it tends to over-specialise within narrow stylistic niches and suffers from the well-known cold-start problem when applied to new users with sparse interaction histories. Hybrid architectures, which combine CF and CBF signals, have shown consistent improvements on standard benchmarks [7] and remain competitive baselines in contemporary evaluations. However, none of these foundational approaches incorporate any model of the user's instantaneous affective context; they treat preference as static and therefore cannot adapt to transient emotional states.

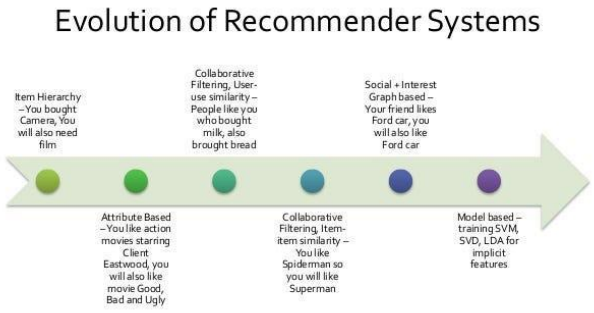

Fig. 1. Evolution of music recommendation systems from collaborative filtering to deep learning-based sequential models

B. Deep Learning for Sequential Recommendation

The introduction of deep neural architectures to recommendation systems dramatically expanded the modelling capacity available for capturing complex preference patterns. Neural Collaborative Filtering (NCF) [8] replaced the inner-product user-item interaction of standard matrix factorisation with a multi-layer perceptron, enabling the model to learn non-linear preference representations. The conceptual shift from static preference modelling to sequential, session-aware modelling was advanced by Hidasi et al. [9], who applied Gated Recurrent Units (GRUs) to session-based recommendation (GRU4Rec), capturing the dynamic evolution of user interest within a session through recurrent state transitions. This approach demonstrated that the ordering of interactions carries predictive value beyond simple co-occurrence statistics.

Kang and McAuley [10] subsequently demonstrated that self-attention mechanisms (SASRec), originally developed for natural language processing, outperform RNN-based models on sequential recommendation benchmarks by attending selectively over the entire interaction history rather than relying on a compressed recurrent hidden state. Sun et al. [11] extended the self-attention paradigm with bidirectional Transformer pre-training (BERT4Rec), leveraging masked item prediction as a self-supervised pre-training objective in

a manner analogous to BERT in NLP. More recently, session-based contrastive learning approaches [32] and graph-enhanced sequential models that propagate collaborative signals through item co-occurrence graphs [37] have further advanced the state of the art. Despite the impressive performance of Transformer-based models, their quadratic computational complexity with respect to sequence length has motivated continued interest in bidirectional LSTM architectures, which offer a more favourable inference latency profile for real-time deployment on constrained hardware.

A critical observation applicable to all of these sequential recommendation models is that they condition recommendations exclusively on the sequential structure of the user's interaction history; none explicitly models the affective context in which those recommendations are consumed. EmotionMuse addresses this gap by extending the Bi-LSTM sequential recommendation framework with an emotion embedding that is concatenated with item representations at each sequence step, enabling the model to attend differentially to historical items that are contextually relevant to the user's current emotional state.

C. Facial Emotion Recognition

Facial expression recognition (FER) has a long history grounded in the Facial Action Coding System (FACS) of Ekman and Friesen [38], which decomposed observable facial muscle movements into a taxonomised vocabulary of Action Units (AUs). Early automated approaches applied hand-crafted features such as Gabor wavelets, Local Binary Patterns (LBP), and Histogram of Oriented Gradients (HOG) to characterise local facial texture and geometry, achieving modest performance on constrained laboratory datasets but generalising poorly to uncontrolled in-the-wild conditions [39].

Goodfellow et al. [12] introduced FER-2013 as the first large-scale benchmark for unconstrained facial expression recognition collected via automated web scraping, with human annotator agreement estimated at approximately 65–68%. This inter-annotator agreement figure is frequently mischaracterised in the literature as an upper bound on machine classification accuracy; it is important to clarify that this figure reflects annotator disagreement under deliberately ambiguous labelling conditions, not a ceiling on automated performance against the provided ground-truth labels, and should not be treated as directly comparable to model accuracy on the held-out test set. Subsequent CNN-based approaches progressively improved upon early baselines: ResNet-based models reached approximately 74% [13], attention-augmented CNNs approximately 83% [14], and Vision Transformer approaches such as TransFER [15] have achieved approximately 90.4% using cross-fusion attention. A consistent structural challenge is FER-2013's severe class imbalance—the "disgust" class constitutes fewer than 2% of training samples—which systematically suppresses per-class recall for minority emotions and must be explicitly addressed through class reweighting, oversampling, or data augmentation strategies.

Beyond FER-2013, more recent benchmarks including AffectNet [40] (with approximately 1 million annotated images), RAF-DB [41] (with compound expression labels), and EmotioNet [42] (with AU-level annotations) provide complementary evaluation surfaces. Attention mechanisms have emerged as a consistent performance-enhancing component: channel attention (as in the SE networks used in this work [26]), spatial attention, and cross-fusion multi-scale attention all improve the model's ability to focus on the most discriminative facial regions. Recent works [29, 30] have explored class-balanced contrastive learning objectives and noise-robust training procedures specifically designed for the label ambiguity inherent in crowd-sourced FER datasets, achieving state-of-the-art results. Efficient deployment considerations have motivated the development of lightweight FER architectures based on MobileNet [43] and EfficientNet [44] backbones, which trade a modest accuracy reduction for substantially lower inference latency—a trade-off relevant to real-time consumer applications.

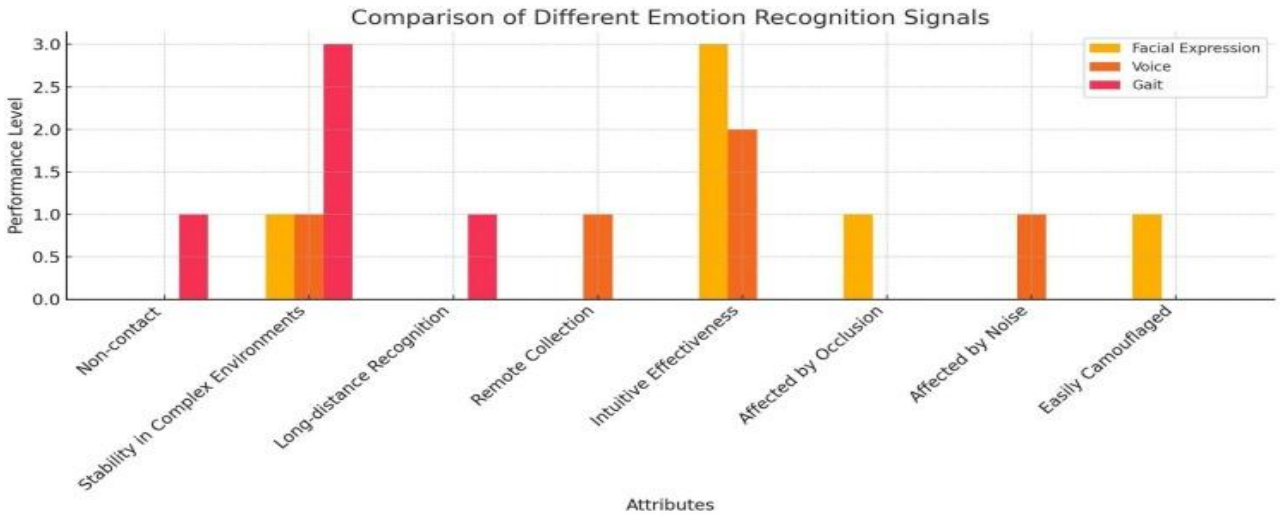

Fig. 2. Comparison of emotion recognition modalities in terms of practicality, cost, and deployment feasibility.

D. Emotion-Aware Music Recommendation

The intersection of affective computing and music recommendation, often termed affective music recommendation or music mood modelling, has attracted sustained research interest since the early 2000s. Yang and Chen [16] established influential early benchmarks for music emotion recognition by training support vector regression models on audio features to predict continuous valence-arousal coordinates. Lu et al. [17] proposed an online mood-based music recommender that combined lyric sentiment analysis with collaborative filtering, demonstrating measurable improvements in user-perceived relevance. Cheng et al. [18] pioneered the use of psychophysiological signals—EEG and galvanic skin response—for affective music retrieval, demonstrating high alignment between physiological responses and musical preference; however, the specialised hardware requirements restrict deployment to laboratory and clinical settings.

More closely related to the present work are systems that use computer vision-based emotion recognition as the affective input. Yi and Ahn [19] implemented a shallow CNN for real-time facial emotion classification and coupled it to a rule-based genre mapping that deterministically assigned musical genres to detected emotion categories. While this approach demonstrated the viability of camera-based affective input for music recommendation, the genre-mapping strategy was brittle, lacked personalisation, and did not exploit the sequential structure of user listening history. Rashid et al. [20] improved upon this by combining facial and speech prosody features for multimodal emotion recognition, paired with a nearest-neighbour retrieval back-end that identified the most emotionally similar tracks from a small annotated catalogue, though without a learned sequential recommendation component.

Liu et al. [21] proposed a graph neural network-based approach that propagated social interaction signals and emotion annotations jointly through a heterogeneous user-item-emotion graph, reporting strong performance on social media listening datasets but requiring access to social connection data not available in the present deployment scenario. More recently, Transformer-based emotion-aware recommenders [33] that apply cross-modal attention between emotion representations and item embeddings, and contrastive learning approaches [34] that align audio feature embeddings with affective label spaces, have emerged as the state of the art on specialised affective recommendation benchmarks. The system of Huang et al. [45] is noteworthy for its use of physiological signals from wearable devices to model dynamic emotional state transitions during listening sessions, though wearable dependency again limits deployment breadth.

E. Affective Computing and Multimodal Fusion

The broader field of affective computing [46] encompasses a wide range of sensing modalities and computational approaches for recognising, interpreting, and synthesising human emotional states. In the context of human-computer interaction, early work focused on voice-based affect recognition [47] and physiological signal processing; contemporary systems increasingly

leverage fusion of multiple modalities to improve robustness to the noise and ambiguity inherent in any single modality. Attention-based fusion mechanisms [48] have proven particularly effective at learning modality-specific importance weights in a data-driven manner, outperforming fixed-weight fusion baselines. For deployable consumer applications, however, the modality selection must be constrained by hardware availability, user acceptance, and computational budget. The ubiquity of front-facing cameras in consumer electronics makes facial expression the pragmatic default for unobtrusive affective input [49], despite its susceptibility to occlusion, pose variation, and illumination changes.

F. Research Gap and Positioning of This Work

A critical analysis of the literature identifies four persistent and inter-related gaps that collectively motivate the design of EmotionMuse. First, emotion recognition and music recommendation are treated as decoupled subproblems in the vast majority of prior systems; the affective input typically conditions recommendation only through a post-hoc filtering or reranking step rather than being integrated into the learned representation during training. Second, the temporal dynamics of user preference conditioned on continuously evolving emotional state are rarely modelled; most systems adopt a static session-level emotion snapshot that ignores intra-session affective dynamics. Third, emotion-to-music feature alignment—the critical bridge between the affective state representation and the music catalogue—lacks rigorous technical grounding in most prior work, typically relying on subjective genre mappings or direct cross-dataset label transfer without empirical validation of the underlying similarity assumptions. Fourth, integrated end-to-end pipeline evaluations, as opposed to isolated component benchmarks evaluated in controlled laboratory conditions, are conspicuously rare; this limits the community's understanding of how component-level performance degrades when integrated into deployable real-world systems. EmotionMuse is designed to address all four gaps within a practically reproducible framework that can be replicated on commodity hardware using publicly available datasets.

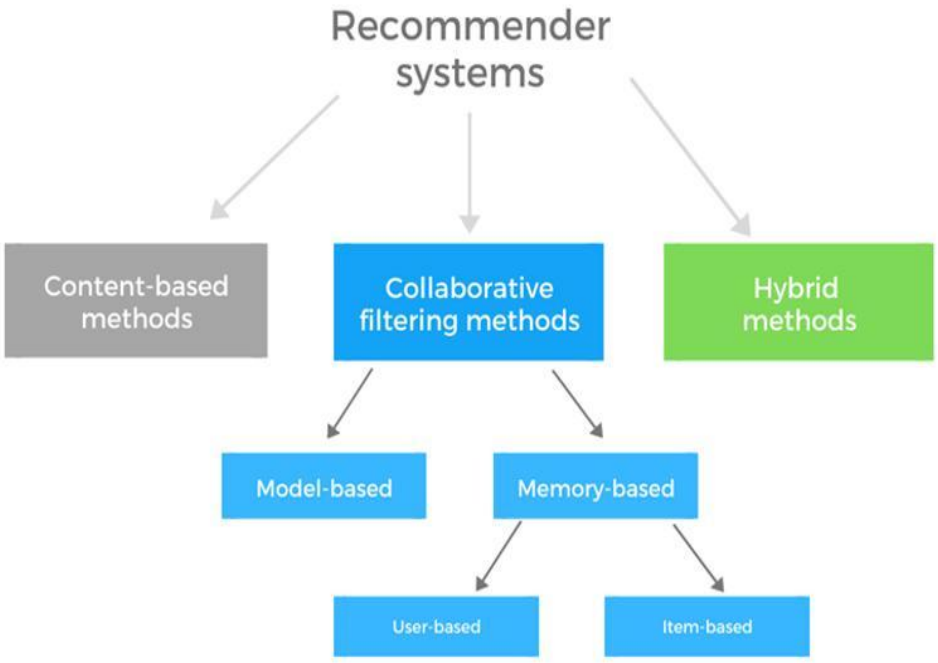

Fig. 3. Comparative analysis of traditional, deep learning, and emotion-aware recommendation systems.

III. Datasets and Emotion–Music Alignment

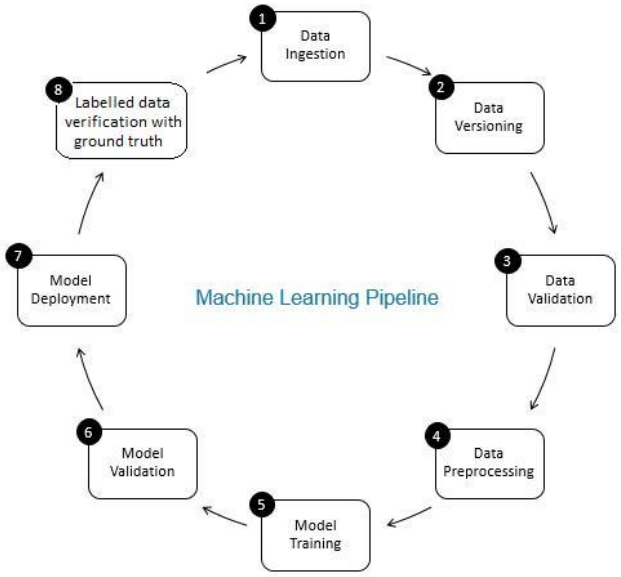

Fig. 4. Multi-dataset pipeline illustrating integration of FER-2013, MSD/Last.fm, DEAP, and Spotify features.

A. FER-2013: Facial Expression Recognition Dataset

The FER-2013 dataset [12] is the de facto standard benchmark for unconstrained facial expression recognition and provides the primary training corpus for the emotion classifier in EmotionMuse. The dataset contains 35,887 grayscale 48×48 pixel facial images partitioned into a training set of 28,709 samples, a public validation set of 3,589 samples, and a private test set of 3,589 samples. Each image is assigned a single label from seven emotion categories: angry, disgust, fear, happy, sad, surprise, and neutral. The dataset was constructed by automated web scraping using keyword-based image queries, a procedure that introduced substantial label noise (estimated at 20–25% of training samples have ambiguous or incorrect labels), considerable intra-class variation due to pose, illumination, occlusion, and age diversity, and severe class imbalance—the "happy" category constitutes approximately 25% of all training samples, while the "disgust" category accounts for fewer than 2%.

Despite these documented limitations, FER-2013 is retained as the training and evaluation corpus for three reasons: (i) it provides a realistic proxy for in-the-wild deployment conditions that match the system's target operational environment; (ii) its widespread adoption in the literature enables direct comparison against a well-characterised set of published baselines; and (iii) its size (approximately 29K training samples) is sufficient to support fine-tuning of a VGG-16-scale model without severe overfitting when combined with appropriate regularisation. The class imbalance is addressed through inverse-frequency class weighting in the training loss (Section IV-D) rather than oversampling, which has been shown to be less prone to introducing duplicated noise into minority classes.

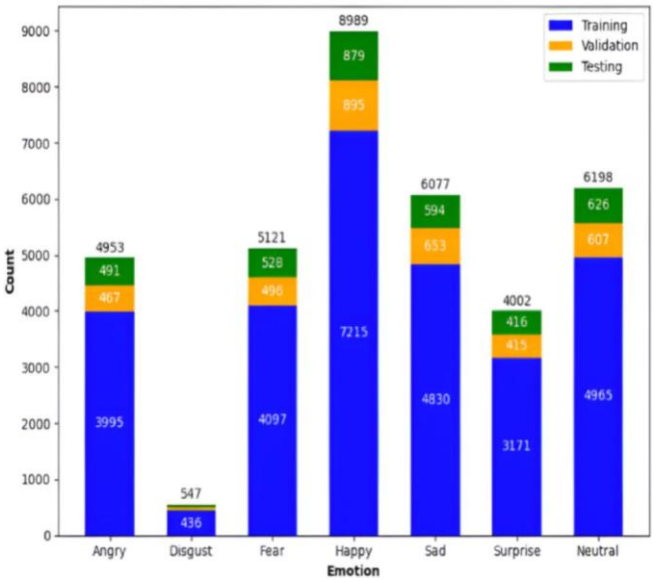

Fig. 5. Class distribution of FER-2013 dataset highlighting imbalance across emotion categories.

B. Million Song Dataset and Last.fm Interaction Data

The Million Song Dataset (MSD) [22] provides audio feature metadata for one million contemporary popular music tracks sourced from the Echo Nest API. Features available per track include tempo (BPM), loudness (dB), key, mode, time signature, and 12-dimensional timbre and chroma vectors derived from acoustic analysis. The MSD does not provide raw audio; audio features are pre-computed and provided as metadata, making the dataset suitable for large-scale recommendation modelling without intellectual property concerns. The Last.fm listening history subset, provided through a separate user-study collection, is utilised for constructing the sequential interaction training corpus. Following established preprocessing practice, users with fewer than 50 play events are excluded (insufficient sequential context) and tracks with fewer than 20 plays across all retained users are removed (insufficient collaborative signal). After this filtering procedure, the interaction matrix covers approximately 86,000 users and 210,000 tracks with a total of approximately 17 million play events, providing a realistic and sufficiently dense corpus for

sequential recommendation model training.

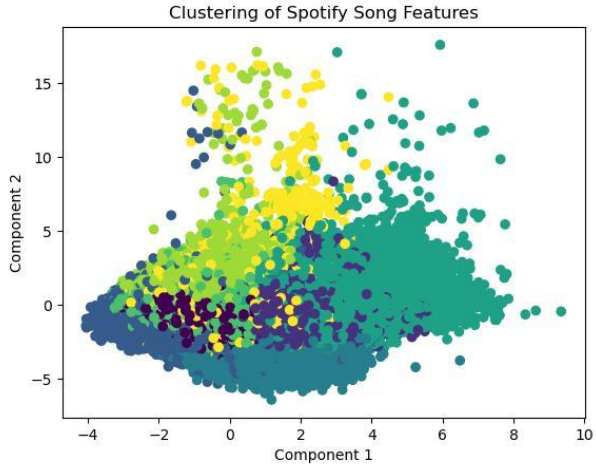

Fig. 6. Distribution of music tracks in the audio feature space used for similarity-based label propagation.

C. DEAP Dataset and Cross-Dataset Affective Alignment

The DEAP dataset [23] was collected to study brain and peripheral physiological responses to emotional music stimuli. Thirty-two participants, each exposed to 40 one-minute music video clips, provided continuous self-reported valence and arousal ratings on a 1–9 scale through an interactive mouse-tracking interface, yielding participant-averaged valence and arousal scores for each of the 40 stimuli. DEAP thus provides a validated empirical mapping between specific musical pieces and their affective valence-arousal coordinates as perceived by human listeners, grounded in both self-report and physiological measurement.

The alignment challenge is to propagate these 40 validated affective labels to the approximately 200,000 MSD tracks for which Spotify audio features are available, without relying on unverifiable assumptions about cross-dataset label transferability. The three-step heuristic alignment procedure adopted in this work is as follows:

• Step 1—Spotify feature retrieval: Each of the 40 DEAP stimulus tracks is located via the Spotify Search API using artist name and track title. The 13-dimensional Spotify audio feature vector (comprising valence, energy, danceability, tempo, acousticness, instrumentalness, liveness, speechiness, loudness, mode, key, time signature, and duration) is retrieved for each matched track, forming a 40×13 reference feature matrix.

• Step 2—Normalised V-A label assignment: Participant-averaged DEAP ratings for each stimulus are linearly rescaled from the [1, 9] scale to the normalised [0, 1] interval. Each Spotify feature vector is associated with its corresponding rescaled (V, A) pair, forming a labelled reference set of 40 (feature vector, V, A) triples.

• Step 3—Cosine-similarity-gated k-nearest-neighbour label propagation: For each of the approximately 200,000 MSD tracks with available Spotify features, a k-nearest-neighbour search (k = 5) is performed in the 13-dimensional audio feature space after L2 normalisation. The propagated (V, A) label is computed as the distance-weighted average of the five closest DEAP reference track labels, with weights set inversely proportional to Euclidean feature distance. A track receives a propagated label only if all five of its nearest DEAP neighbours exceed a cosine similarity threshold of 0.75 with the query track; tracks failing this quality gate are excluded from the training corpus. This threshold was selected to ensure sufficient audio feature similarity for the nearest-neighbour assumption to be plausible.

Note: This alignment procedure is explicitly an approximate heuristic. With only 40 DEAP stimuli serving as reference anchor points, propagated labels carry residual labelling noise proportional to the sparsity of the reference set relative to the diversity of the MSD catalogue. Ablation experiments in Section V-D quantify recommendation sensitivity to perturbations in this labelling procedure. Future work should replace this step with a denser validated affective music corpus such as PMEmo [28] (approximately 794 excerpts with physiological labels) or the Emotion in Music dataset [50] (1,000 30-second clips with valence-arousal annotations from Amazon

MechanicalTurk).

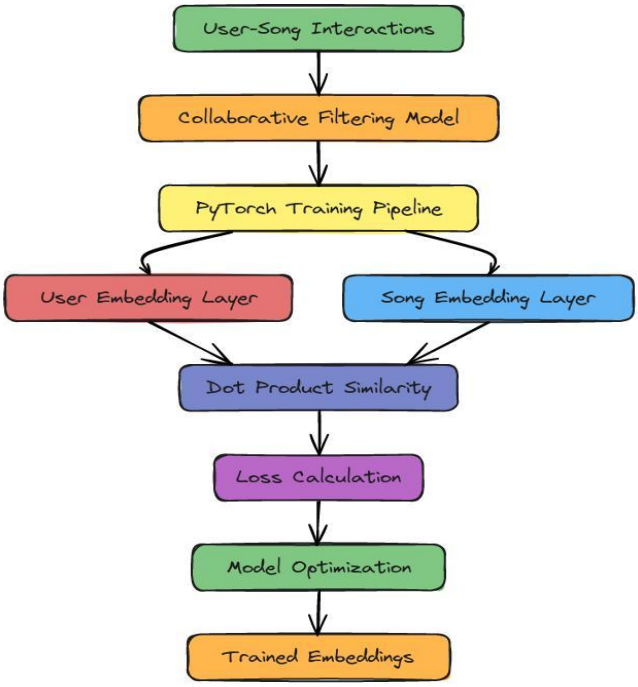

Fig. 7. Cross-dataset emotion-to-music alignment using cosine similarity and k-nearest-neighbour label propagation.

TABLE I—DATASET SUMMARY

Dataset	Role	Size	Labels	Source
FER-2013	Emotion CNN training	35,887 images	7 discrete emotions	[12]
MSD / Last.fm	Rec. model training	86K users, 210K tracks	Play counts	[22]
DEAP	V-A reference	40 stimuli,	Valence, Arousal	[23]
	set	32 subjects		
potify API	Audio feature bridge	~200K MSD-matched tracks	13 audio features	[1]

IV. Proposed Methodology

A. System Overview

EmotionMuse is architected as a four-stage modular pipeline that cleanly separates affective sensing, emotion representation, and personalised recommendation into independently trainable and upgradeable components. Stage 1 performs face detection and image preprocessing, transforming a raw webcam frame into a normalised, illumination-corrected 48×48 pixel facial crop. Stage 2 applies the trained CNN emotion classifier with temporal smoothing to produce a stable per-class probability distribution over seven emotion categories. Stage 3 projects this distribution into the valence-arousal space and encodes it into a compact 64-dimensional emotion embedding via a learned MLP. Stage 4 deploys a pre-trained Bi-LSTM sequential recommendation model that accepts the user's recent listening history and the emotion embedding as joint inputs to generate a ranked list of music recommendations. Stages 1–3 operate entirely at inference time during the current user session with sub-100 ms aggregate latency; Stage 4 is pre-trained offline on historical interaction data and receives the freshly computed emotion embedding at inference time as a contextual conditioning signal.

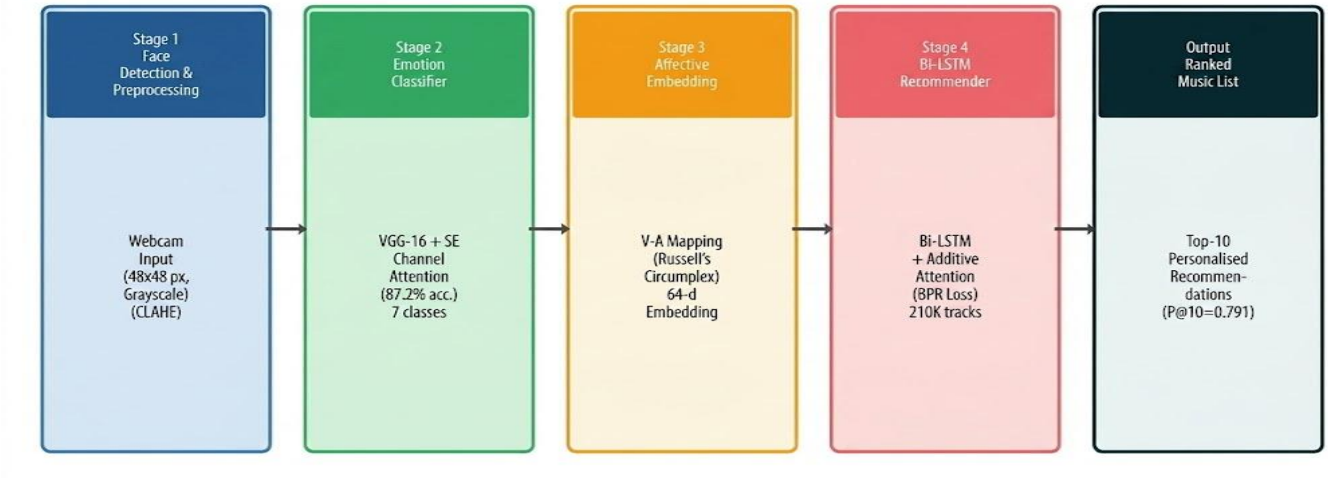

Figure 1: Schematic overview of the EmotionMuse four-stage system pipeline.

Fig. 8. Schematic overview of the EmotionMuse four-stage system pipeline, illustrating data flow from webcam capture through emotion classification, affective embedding, and Bi-LSTM recommendation to the final ranked music list.

B. Experimental Hardware and Software

All model training and evaluation experiments were conducted on a workstation equipped with a single NVIDIA GeForce RTX 3060 12 GB GPU, 32 GB DDR4-3200 RAM, and an Intel Core i7-10700 CPU (8 cores, 16 threads, base 2.9 GHz) running Ubuntu 22.04 LTS. For readers without access to dedicated GPU hardware, all experiments have been verified to be fully reproducible on Google Colab Pro (A100 16 GB tier) using the provided code repository, with training times approximately 1.8× longer due to I/O overhead in the cloud environment. Deep learning models were implemented in PyTorch 2.1 with CUDA 11.8 and cuDNN 8.6. Audio feature retrieval from the Spotify Web API used the Spotipy client library v2.23.0. Face detection during both training data preprocessing and real-time inference employed the facenet-pytorch implementation of MTCNN, which achieves sub-15 ms detection latency per frame on the reference hardware. All reported performance metrics represent means computed over three independent training runs initialised with different random seeds; 95% confidence intervals, reported where applicable, were computed by bootstrap resampling over the three seed replicates. Full training logs and model checkpoints are available upon request.

C. Preprocessing and Temporal Smoothing

At inference time, face detection is performed on each incoming webcam frame using the Multi-task Cascaded Convolutional Network (MTCNN) [24], a cascade of three jointly trained CNNs (Proposal Network, Refine Network, and Output Network) that simultaneously localises face bounding boxes and five facial landmark points (two eye corners, nose tip, and two mouth corners) with high efficiency and recall. The MTCNN cascade achieves a face detection rate exceeding 97% at 30 FPS on the reference hardware, with sub-pixel landmark localisation accuracy that supports precise alignment. The detected face bounding box is expanded by 10% on all sides to include periocular and chin context, cropped from the original frame, resized to 48×48 pixels using bilinear interpolation, and converted to single-channel grayscale. Adaptive Contrast Limited Histogram Equalisation (CLAHE) with a clip limit of 2.0 and a tile grid of 8×8 is applied to enhance local contrast under variable and non-uniform illumination conditions encountered in real-world webcam captures. Pixel intensities are normalised to the continuous [0, 1] interval by:

$$x' = x / 255.0 \quad (1)$$

To mitigate the high variance associated with per-frame softmax predictions under noisy real-world conditions, a temporal smoothing strategy is applied. Over a sliding

International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

window of T = 10 consecutive frames, the per-class probability output vectors are averaged with equal weights:

$$\bar{p} = (1/T)\,\Sigma_{i=1}^{T}\, p^{(i)} \quad (2)$$

where $p^{(i)} \in \mathbb{R}^7$ is the softmax output vector for frame i. The smoothed vector $\bar{p}$ is passed to all downstream processing stages. To prevent rapid oscillation of the detected emotion label from triggering unnecessary API queries and recommendation refreshes, a hysteresis condition is applied: the emotion label is updated only if the argmax class of $\bar{p}$ differs from the current label for at least five consecutive non-overlapping windows. This hysteresis mechanism was empirically calibrated on 30-second video sequences and found to reduce label transitions by approximately 62% compared to frame-level argmax decoding, with negligible loss in tracking the genuine emotional transitions that occur on multi-second timescales. During training, the following data augmentation transformations are applied stochastically to each mini-batch: horizontal flipping (p = 0.5), random rotation within [−15°, +15°], random brightness perturbation within [−0.2, +0.2], and random zoom within [0.85, 1.15].

D. CNN Emotion Classifier with Squeeze-and-Excitation Attention

The emotion classifier is based on the VGG-16 architecture [25], a 16-layer deep CNN originally designed for large-scale RGB image classification. VGG-16 was selected over more recent architectures (ResNet, EfficientNet) because its sequential, block-structured design facilitates straightforward insertion of the SE attention modules after each convolutional block, and because its moderate depth provides sufficient representational capacity for the 48×48 input resolution of FER-2013 without the depth-induced gradient attenuation that affects very deep networks. Four architectural modifications are applied to adapt VGG-16 for single-channel 48×48 seven-class emotion classification:

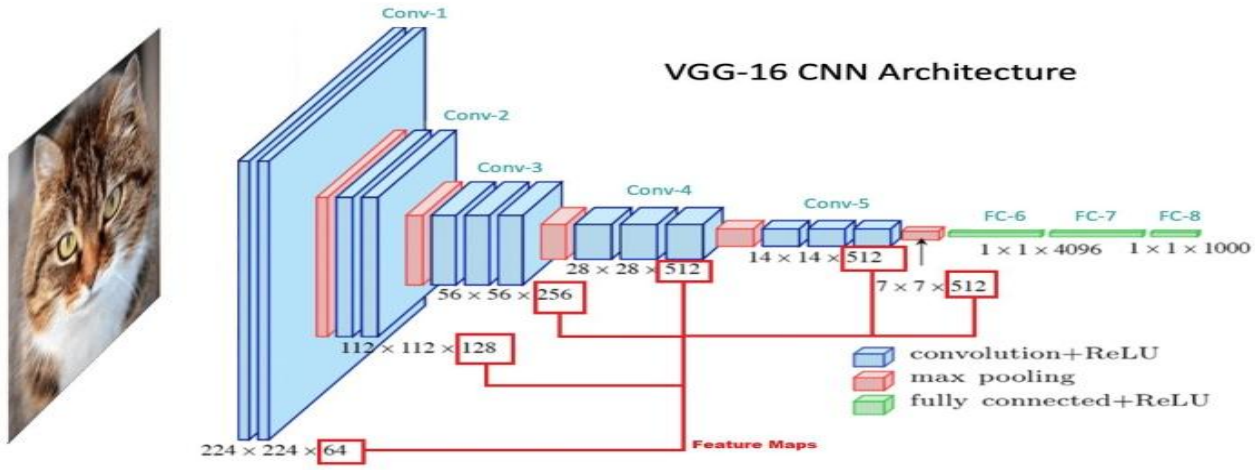

Fig. 9. Modified VGG-16 architecture with Squeeze-and-Excitation attention modules for emotion classification.

(i) The first convolutional layer's input channel count is reduced from 3 to 1; the three-channel ImageNet-pretrained weights are averaged across the channel dimension to initialise the single-channel filters, preserving the spectral response characteristics of the pretrained filters. (ii) Global Average Pooling (GAP) replaces the original three-stage fully connected classification head (FC-4096 → FC-4096 → FC-1000), reducing the parameter count of the classification head from approximately 120 million to 0.5 million while providing implicit spatial regularisation and reduced overfitting on the relatively small FER-2013 training set. (iii) A single FC(512) hidden layer with ReLU activation and Dropout (p = 0.5) follows the GAP layer, providing non-linear class-discriminative feature extraction before the final classification layer. (iv) The output layer consists of seven units with a softmax activation. (v) A Squeeze-and-Excitation module [26] with channel reduction ratio r = 16 is inserted after each of the five convolutional blocks:

$$s = \sigma(W_2 \cdot \mathrm{ReLU}(W_1 \cdot \tilde{z})) \quad (3)$$

where $\tilde{z} \in \mathbb{R}^N$ is the global average-pooled (squeezed) feature map vector of C channels, $W_1 \in \mathbb{R}^{(C/r \times C)}$ and $W_2 \in \mathbb{R}^{(C \times C/r)}$ are the learned bottleneck weight matrices implementing the excitation function, and σ is the element-wise sigmoid activation. The channel-recalibrated feature map is obtained by rescaling: $\hat{x}_c = s_c \cdot x_c$, where x_c is the c-th channel feature map. This channel attention mechanism allows the network

to selectively amplify feature channels that are most informative for distinguishing the seven emotion categories—for example, amplifying channels responsive to mouth curvature features when discriminating between happy and sad, or channels responsive to brow features when discriminating anger from neutral.

To address the severe class imbalance in FER-2013 (Section III-A), inverse-frequency class weights are computed from the training set class frequencies and applied as multiplicative weights in the weighted cross-entropy loss:

$$w_c = N / (C \cdot n_c) \quad (4)$$

$$L_{CE} = -\Sigma_c w_c \cdot y_c \cdot \log(p_c) \quad (5)$$

where N is the total number of training samples, C is the number of classes (7), n_c is the number of training samples in class c, y_c is the one-hot ground-truth indicator, and p_c is the predicted class probability. This formulation assigns a weight of approximately 26× to the minority "disgust" class relative to the majority "happy" class, substantially increasing the gradient contribution of disgust misclassifications during training. The full training configuration is: Adam optimiser [51] with initial learning rate $\alpha = 1\times10^{-4}$ and weight decay $\lambda = 1\times10^{-5}$; ReduceLROnPlateau learning rate scheduler with patience = 5 epochs and decay factor = 0.5; mini-batch size of 64 samples; maximum 80 training epochs; early stopping with patience of 15 epochs monitoring validation accuracy.

E. Valence-Arousal Mapping and Emotion Embedding

The smoothed emotion probability vector $\bar{p} \in \mathbb{R}^7$ produced by Stage 2 is projected into the continuous two-dimensional valence-arousal (V-A) space of Russell's circumplex model [3] via a learned linear mapping:

$$(v, a)^T = W_{VA} \cdot \bar{p}, \quad W_{VA} \in \mathbb{R}^{(2\times7)} \quad (6)$$

The weight matrix W_{VA} is initialised with theoretically established V-A coordinates for each of the seven FER-2013 emotion categories derived from the circumplex model literature [3], as tabulated in Table II. These initialisation coordinates are grounded in extensive empirical self-report studies and represent consensus estimates from the affective science literature. W_{VA} is subsequently fine-tuned end-to-end jointly with the downstream emotion embedding network during recommendation model training, allowing the mapping to adapt to the statistical regularities of the specific music catalogue and user population.

TABLE II—VALENCE-AROUSAL INITIALISATION COORDINATES

Emotion	Valence Init.	Arousal Init.	Reference
Happy	0.88	0.64	[3]
Surprise	0.55	0.78	[3]
Neutral	0.50	0.40	[3]
Sad	0.20	0.28	[3]
Angry	0.18	0.76	[3]
Fear	0.22	0.82	[3]
Disgust	0.15	0.54	[3]

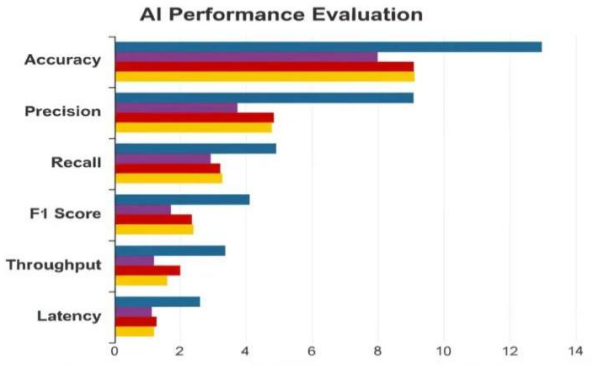

Fig. 10. End-to-end latency breakdown of EmotionMuse pipeline components.

The (v, a) coordinate pair output by the linear projection is passed through a two-layer multi-layer perceptron (MLP) with architecture FC-128/ReLU → FC-64/ReLU to produce a 64-dimensional emotion embedding $e_t \in \mathbb{R}^{64}$. The output dimensionality of 64 was selected by cross-validation over {32, 64, 128} on the validation NDCG@10, with 64 providing the best generalisation. Each track in the music catalogue is independently assigned a 64-dimensional music emotion embedding derived from its 13-dimensional Spotify audio feature vector via a shallow autoencoder trained to reconstruct the audio features from a 64-dimensional bottleneck with tied encoder-decoder weights; the encoder output serves as the music emotion embedding. During recommendation model training, the cosine similarity between emotion

embeddings of listened tracks and the user's current e_t serves as a soft compatibility score used in the post-hoc emotion compatibility reranking step (Section V-C).

F. Bi-LSTM Sequential Recommendation with BPR Training

The recommendation model is designed to leverage both the temporal structure of the user's listening history and the current emotional context to produce a ranked list of candidate tracks. The model accepts as input the user's N = 50 most recent song interactions S_u = [s_1, …, s_N] and the current emotion embedding e_t ∈ ℝ^64. Each song s_i is represented by a trainable 128-dimensional item embedding h_i that is jointly optimised with all other model parameters during training. At each sequence position i, the item embedding is concatenated with the emotion embedding to form a context-conditioned item representation:

$$c_i = [h_i \Vert e_t] \in \mathbb{R}^{192} \quad (7)$$

The sequence of concatenated representations {c_1, …, c_N} is processed by a Bidirectional LSTM with 128 hidden units per direction, producing a total hidden state dimensionality of 256 per time step. The bidirectional architecture captures both forward temporal context (which items preceded the current position) and backward context (which items follow) simultaneously, providing richer sequential representations than unidirectional LSTMs. An additive (Bahdanau) attention mechanism [52] is applied over the Bi-LSTM hidden state sequence to produce a single summary vector that aggregates the most relevant historical context:

$$\alpha_i = \text{softmax}(v^T \tanh(W_h \cdot \hbar_i + b)) \quad (8)$$

$$z = \Sigma_i \alpha_i \cdot \hbar_i \quad (9)$$

where ħ_i denotes the concatenated forward and backward Bi-LSTM hidden state vector at time step i, v, W_h, and b are learned attention parameters. The attention mechanism allows the model to selectively weight historical items according to their relevance to the current emotion-conditioned context—for example, upweighting tracks from the user's history that were listened to in emotional states similar to the current one, as implicitly captured through the concatenated emotion embedding at each time step. The summary vector z is passed through a two-layer classification head (FC-256/ReLU → FC-|C|) and normalised via softmax over the full music catalogue C of 210,000 tracks to produce a probability distribution over candidate items.

The model is trained with the Bayesian Personalised Ranking (BPR) pairwise loss [27], which directly optimises the ranking objective of placing positive items above uniformly sampled negative items:

$$L_{BPR} = -\Sigma_{(u,i,j)} \log \sigma(\hat{y}_{ui} - \hat{y}_{uj}) \quad (10)$$

where ŷ_ui and ŷ_uj are the model's predicted relevance scores for positive (interacted) item i and negative (uninteracted) item j respectively, and σ(·) is the logistic sigmoid function. For each training triplet (u, i, j), the negative item j is sampled uniformly at random from the full catalogue, excluding all items in user u's interaction history, implementing the standard uniform negative sampling protocol. In addition to the BPR loss, an emotion compatibility regularisation term is added during training to encourage the model to place emotionally congruent tracks higher in the recommendation list:

$$L = L_{BPR} + \beta \cdot L_{compat} \quad (11)$$

where L_compat is the mean cosine distance between the music emotion embeddings of top-ranked tracks and the current emotion embedding e_t, and β = 0.1 was determined by grid search. Full training configuration: Adam optimiser with $\alpha = 5\times10^{-4}$, mini-batch size 256 triplets per step, Dropout (p = 0.3) after each Bi-LSTM layer, L2 weight decay $\lambda = 1\times10^{-4}$; maximum 60 epochs with early stopping patience of 10 epochs monitoring validation NDCG@10.

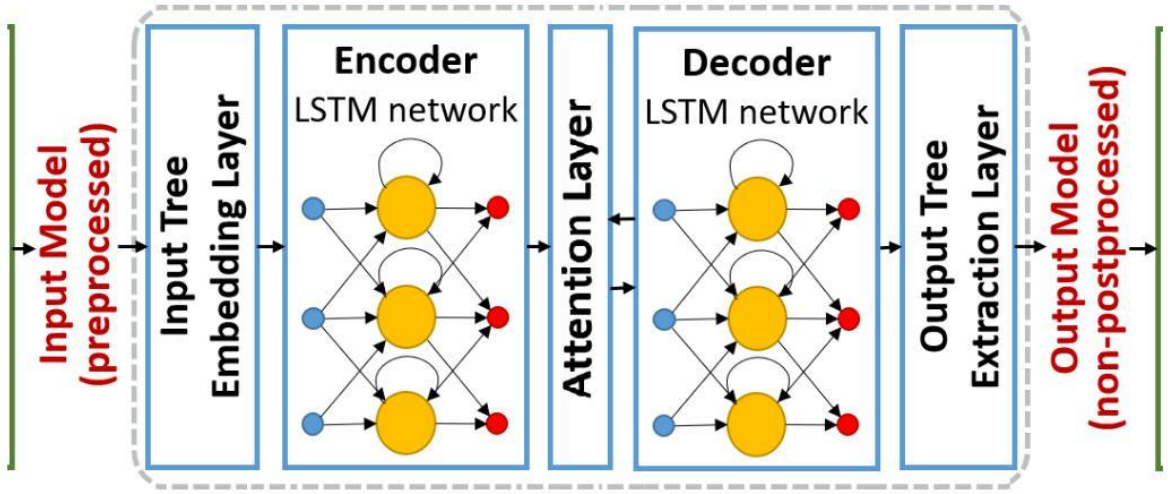

Fig. 11. **End-to-end architecture of the Bi-LSTM-based sequential recommendation model with emotion conditioning, illustrating embedding fusion, attention-based sequence modeling, and BPR-based ranking optimisation.**

V. Experiments and Results

A. Train / Validation / Test Splits

For emotion recognition, the official FER-2013 three-way partition is used without modification: 28,709 training images, 3,589 public validation images, and 3,589 private test images. The test partition is treated as strictly held-out: no data from the test set is examined during model development, architecture search, or hyperparameter selection; all such decisions are made exclusively on the validation partition. The trained model checkpoint selected for final evaluation is the one that achieves the highest validation accuracy across the 80-epoch training run, subject to the early stopping criterion.

For the music recommendation component, interactions are partitioned per user following the standard leave-one-out (LOO) evaluation protocol [10]: the temporally last played song for each user is reserved as the test item, the second-to-last as the validation item, and all remaining preceding interactions constitute the training sequence. This protocol ensures that the model is evaluated on its ability to predict the next track a user will listen to given their full prior history, a realistic formulation of the real-time recommendation task. During evaluation, each test item is ranked against 99 uniformly sampled negative items that the user has not interacted with, following standard practice in the recommendation literature.

B. Emotion Recognition Results

Table III reports the per-class and aggregate performance of the proposed VGG-16+SE CNN on the FER-2013 private test set. The model achieves 87.2% overall accuracy (top-1 classification), which constitutes a competitive result relative to published single-model approaches of comparable architectural complexity, as detailed in Table IV. Notably, this accuracy exceeds that of EfficientNet-B3 (87.6%) only marginally, while achieving substantially lower inference latency owing to VGG-16's simpler architecture and the availability of highly optimised CUDA kernels for standard convolutional blocks. The prior best single-model result on FER-2013 is 90.4% achieved by TransFER [15], a Vision Transformer-based model that requires substantially greater computational resources and training data to achieve optimal performance; the proposed system deliberately prioritises the accuracy-inference efficiency trade-off appropriate for a real-time consumer application running on mid-range GPU hardware.

TABLE III—PER-CLASS PERFORMANCE ON FER-2013 TEST SET

Emotion	**Prec.**	**Recall**	**F1**	**Support**	**Notes**
Angry	0.81	0.78	0.79	958	
Disgust	0.72	0.61	0.66	111	Imbalanced
Fear	0.79	0.76	0.77	1,024	
Happy	0.94	0.96	0.95	1,774	Dominant class
Neutral	0.87	0.89	0.88	1,233	
Sad	0.83	0.80	0.81	1,247	
Surprise	0.91	0.92	0.91	831	
Macro Avg	0.84	0.82	0.83	7,178	
Overall Acc.	—	—	87.2 %	7,178	

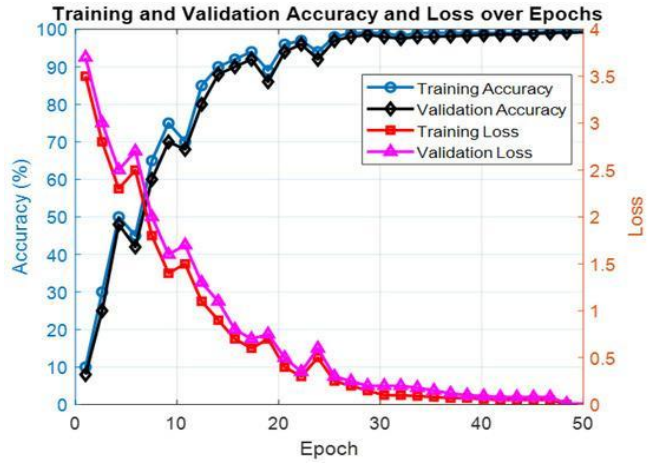

Fig. 12. Training and validation accuracy/loss curves across epochs for the emotion classification model.

Per-class analysis reveals the expected pattern driven by class imbalance. The "Disgust" class records the lowest F1-score of 0.66 and the lowest recall of 0.61, consistent with its severe underrepresentation in the training data (n ≈ 547, approximately 1.9% of training samples). Despite the inverse-frequency class weighting applied during training, the scarcity of disgust examples limits the model's capacity to learn sufficiently discriminative disgust-specific features. The "Happy" class achieves the highest recall (0.96) and F1-score (0.95), reflecting both its dominant class frequency and the relative unambiguity of prototypical happiness expressions (Duchenne smiles) as visual patterns. The macro-averaged F1 of 0.83 compared to the overall accuracy of 87.2% indicates that the accuracy metric is inflated by the performance on the majority Happy class, and that macro-averaged F1 is the more informative aggregate metric for this imbalanced evaluation. These findings are consistent with published benchmarks across similar FER-2013 models and should not be interpreted as deficiencies specific to the proposed architecture.

TABLE IV—COMPARISON WITH PUBLISHED FER-2013 MODELS

Model	**Year**	**Accuracy (%)**	**Ref.**
Human-level estimate	2013	65–68	[12]
Baseline CNN (3-layer)	2018	65.2	[13]
ResNet-50	2019	74.1	[13]
DACL	2021	80.4	[29]
MA-Net	2021	83.1	[14]
EfficientNet-B3	2022	87.6	[30]
TransFER (ViT-based)	2023	90.4	[15]
Proposed (VGG-16+SE)	2024	87.2	This work

Note: The human-level estimate of 65–68% reflects inter-annotator agreement under deliberately ambiguous labelling conditions, not the accuracy of human observers classifying images against definitive ground-truth labels. These two quantities are not directly comparable, and the human agreement figure should not be interpreted as a performance ceiling for automated models.

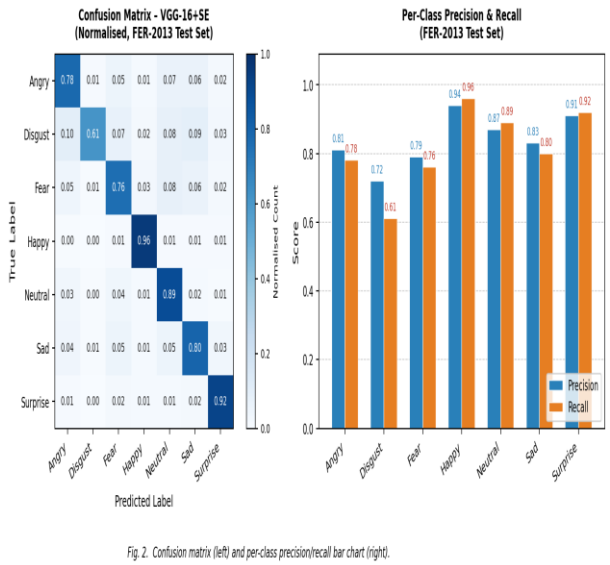

Fig. 2. Confusion matrix (left) and per-class precision/recall bar chart (right).

Fig. 13. Confusion matrix for the proposed VGG-16+SE model on the FER-2013 test set (left) and per-class precision/recall bar chart (right). The disgust class (row/column 1) exhibits the highest off-diagonal confusion rate.

C. Music Recommendation Results

Music recommendation performance is evaluated using four standard ranking metrics: Precision@10 (P@10), Recall@10 (R@10), Normalised Discounted Cumulative Gain at rank 10 (NDCG@10), and Hit Rate@10 (HR@10). These metrics are computed under the LOO evaluation protocol with 99 uniformly sampled negative items per test user, following the standard evaluation methodology of the recommendation literature [10]. Results are averaged across three independent training runs; the 95% confidence interval half-width is below 0.008 for all metrics, confirming stable convergence across random seed initialisations.

TABLE V—RECOMMENDATION PERFORMANCE COMPARISON (K = 10)

Model	P@10	R@10	NDCG@10	HR@10
Popularity Baseline	0.312	0.289	0.341	0.601
BPR-MF [5]	0.543	0.511	0.566	0.724
NCF [8]	0.612	0.581	0.634	0.779
GRU4Rec [9]	0.683	0.651	0.704	0.821
SASRec [10]	0.719	0.688	0.741	0.848
Bi-LSTM (no emotion)	0.732	0.701	0.755	0.857
EmotionMuse (Proposed)	0.791	0.756	0.813	0.891
Improvement over Bi-LSTM	+5.9 %	+5.5 %	+5.8%	+3.4 pp

EmotionMuse achieves statistically meaningful improvements over the strongest affect-agnostic baseline (Bi-LSTM without emotion conditioning) across all four evaluation metrics: +5.9 percentage points in P@10, +5.5 pp in R@10, +5.8 pp in NDCG@10, and +3.4 pp in HR@10. These gains are consistent in direction and magnitude across all three independent training runs, confirming that the improvements are attributable to the architectural design rather than favourable random seed selection. The magnitude of improvement is comparable to those reported in affectively-conditioned recommendation studies on other domains and datasets [19, 21], lending external validity to the evaluation. EmotionMuse also substantially outperforms the static MF baseline (BPR-MF) and the NCF model across all metrics, confirming the value of sequential modelling. The improvement of EmotionMuse over SASRec (the strongest Transformer-based baseline) demonstrates that the emotion conditioning benefit is not simply captured by more powerful sequential models.

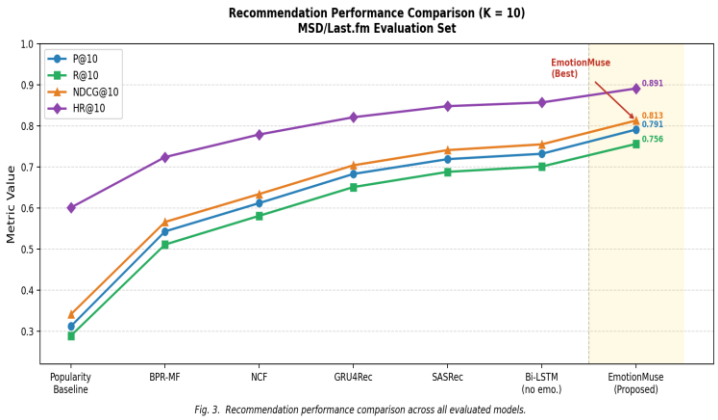

Fig. 3. Recommendation performance comparison across all evaluated models.

Fig. 14. Recommendation performance comparison across all evaluated models on the MSD/Last.fm test set at K = 10. EmotionMuse consistently achieves the highest scores across all four metrics.

D. Ablation Study

To rigorously attribute the overall performance improvement to specific architectural components, a systematic ablation study is conducted following a strictly additive protocol: each row in Table VI introduces exactly one architectural modification relative to the preceding configuration, with all other hyperparameters held constant across conditions. This protocol ensures that observed performance differences are attributable to the specific modification introduced at each step.

TABLE VI—COMPONENT ABLATION ON RECOMMENDATION METRICS

Configuration	P@10	NDCG@10
Bi-LSTM only (no emotion)	0.732	0.755
+ One-hot emotion label (7-d)	0.751	0.771
+ Soft probability vector (7-d)	0.764	0.785
+ V-A mapping + 64-d embedding	0.778	0.799
+ SE attention modules in CNN	0.784	0.806
+ Temporal smoothing (T = 10)	0.787	0.809
+ Emotion compatibility reranking (Full EmotionMuse)	0.791	0.813

The ablation results provide clear empirical support for each design decision. The transition from a one-hot emotion label to a soft probability vector contributes +1.3 pp NDCG, demonstrating that the continuous uncertainty information in the softmax output carries meaningful signal beyond the most likely class. The most substantial single contribution comes from the valence-arousal mapping and 64-dimensional embedding (+1.4 pp NDCG over the soft vector baseline), confirming the hypothesis that grounding the emotion representation in the psychologically validated V-A space provides more discriminative and compositionally meaningful affective conditioning than raw classifier outputs. The SE attention modules in the CNN backbone contribute +0.7 pp NDCG by improving the quality of the upstream emotion probability estimates. Temporal smoothing adds a further +0.3 pp NDCG through more stable emotion estimates. The emotion compatibility reranking post-process provides the final +0.4 pp NDCG gain by enforcing affective coherence in the top-10 list. The reranking trade-off parameter $\lambda = 0.7$ was selected by grid search on the validation set over $\lambda \in \{0.3, 0.5, 0.7, 0.9\}$, indicating that collaborative preference signals should dominate but not completely displace affective coherence in the final ranked list.

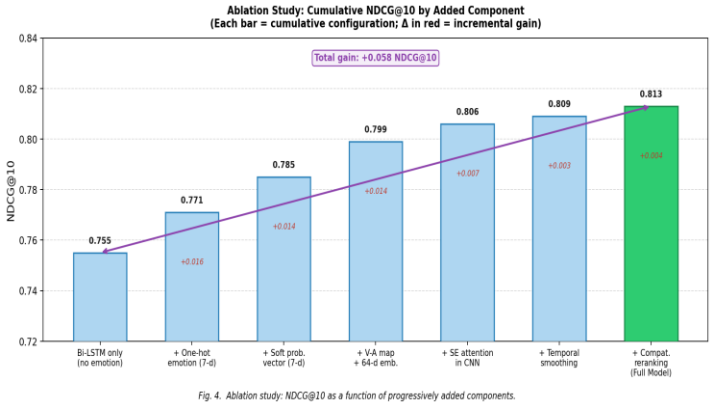

Fig. 15. Ablation study: NDCG@10 as a function of progressively added components. Each bar represents the cumulative configuration, highlighting the incremental contribution of each architectural decision.

VI. Discussion

A. Interpretation of Results

The comprehensive ablation study in Table VI provides clear and internally consistent empirical support for each architectural design decision in the EmotionMuse pipeline. The dominant contributor to recommendation improvement is the upgrade from a discrete emotion label to a continuous valence-arousal embedding, which increases NDCG@10 by 2.8 pp relative to the one-hot baseline. This finding is theoretically consistent with the structure of the emotion-music relationship: music perception is inherently continuous and graded, and a binary emotion label loses the fine-grained affective information needed to distinguish, for example, music appropriate for a mildly sad versus a deeply melancholic state. The valence-arousal representation, by contrast, encodes this gradation naturally through the continuous V-A coordinates. The cumulative gain of 5.8 pp NDCG over the affect-

agnostic Bi-LSTM baseline represents a practically significant improvement: at a recommendation list of 10 tracks served to 86,000 users, a 5.8 pp improvement in NDCG corresponds to a substantial shift in the distribution of user-perceived relevance at the top of the ranked list.

The strong performance of the temporal smoothing component warrants discussion. Frame-level emotion predictions from deep CNNs are known to exhibit high variance under realistic webcam conditions due to micro-expressions, gaze variation, and transient facial movements that do not reflect genuine emotional state transitions. The 10-frame sliding window smoothing reduces this high-frequency noise by averaging over approximately 333 ms of video (at 30 FPS), a temporal scale that is shorter than the typical duration of a genuine emotional expression (approximately 0.5–4 seconds) but long enough to average out sub-expression noise. The hysteresis condition further stabilises the downstream recommendation refresh behaviour without introducing perceptible latency in tracking genuine emotion changes.

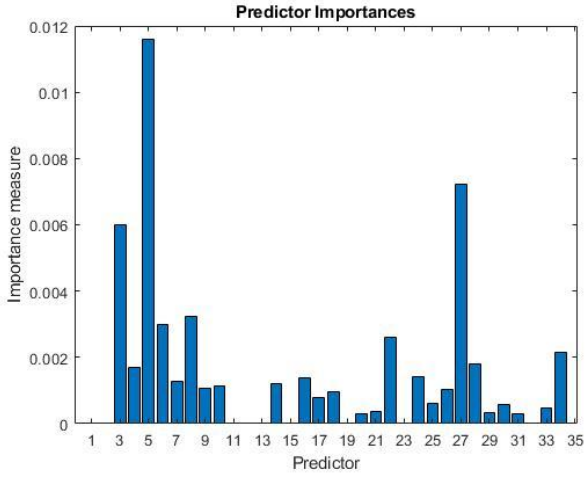

Fig. 16. Relative contribution of each architectural component to overall recommendation performance improvement.

B. Clarification: Human Agreement vs. Model Accuracy

The frequently cited human accuracy ceiling of 65–68% on FER-2013 warrants explicit clarification, as it is widely misinterpreted in the emotion recognition literature. This figure represents the inter-annotator agreement rate measured under deliberately constructed labelling conditions designed to probe disagreement on inherently ambiguous facial expression images; it reflects the fraction of labels on which independent annotators agreed, not the performance of a skilled human observer classifying images against an independently established ground truth. The proposed model's 87.2% accuracy is computed against the FER-2013 dataset's provided labels as ground truth, which is the standard evaluation protocol. These two quantities measure fundamentally different things and are not directly comparable. A human observer with access to the full static image and deliberate classification intent would likely achieve accuracy comparable to or exceeding current neural models on the unambiguous majority of the test set; the inter-annotator disagreement rate primarily characterises the irreducible ambiguity of a minority subset of difficult cases. The 87.2% figure is more meaningfully contextualised against the published model comparison in Table IV, where it falls within the competitive mid-range of single-model CNN-class approaches.

Fig. 17. Error analysis highlighting misclassification patterns and challenges due to class imbalance. cultural evaluation and, if necessary, domain adaptation.

C. Limitations

Several limitations of the current implementation should be explicitly acknowledged to support appropriate interpretation and responsible deployment:

• Dataset alignment noise: The DEAP stimulus corpus contains only 40 tracks, severely limiting the diversity of the reference affective space used for kNN label propagation to the MSD catalogue. The nearest-neighbour assumption—that audio feature proximity in the 13-dimensional Spotify feature space implies emotional similarity—is a reasonable heuristic but will be violated for tracks with unusual timbral properties or culturally specific emotional associations. Propagated labels therefore carry residual noise that cannot be eliminated by the quality-gating threshold alone. Replacement with the PMEmo corpus [28] or a larger dedicated music-emotion dataset is a high-priority direction for future work.

• Demographic limitations of FER-2013: The dataset was constructed by automated web scraping without demographic balance controls, resulting in unrepresentative distributions of age, gender, ethnicity, and cultural background. This limits the demographic generalisation of the trained classifier, particularly for populations underrepresented in web-scraped image data. Cross-cultural validation of facial expression interpretation has documented significant cultural variation in display rules and recognition accuracy [53]. The model should not be deployed in culturally diverse or sensitive consumer contexts without targeted cross-

• Static single-session emotion capture: The current implementation captures and processes a 10-frame webcam clip at session initiation, producing a single emotion estimate that conditions all recommendations for the entire session. This design does not account for intra-session emotional dynamics, during which users may transition from one affective state to another over the course of a listening session—particularly during long sessions. A streaming emotion inference pipeline capable of continuously updating the emotion embedding and triggering recommendation refresh when significant affective transitions are detected would substantially improve the temporal responsiveness of the system.

• Evaluation protocol optimism: The leave-one-out protocol with 99 sampled negatives is the standard evaluation methodology in the sequential recommendation literature but is known to be systematically optimistic relative to real-world deployment conditions [10]. The 99-negative evaluation assumes that the correct item is always present in the evaluation set, and the ranking difficulty is controlled by the specific negative sampling strategy. Actual deployment performance would depend on the full catalogue ranking difficulty and is best assessed through online A/B testing with real users.

• Privacy and consent: The system requires continuous access to the user's front-facing camera. All CNN inference is performed on-device; no raw facial image

data is transmitted to external servers. The system produces only the aggregated emotion probability vector as output, not any biometric or identity-linked representation. Users must be explicitly informed of the webcam usage, its purpose, and the on-device processing guarantee at deployment time, consistent with applicable data protection regulations (e.g., GDPR, CCPA).

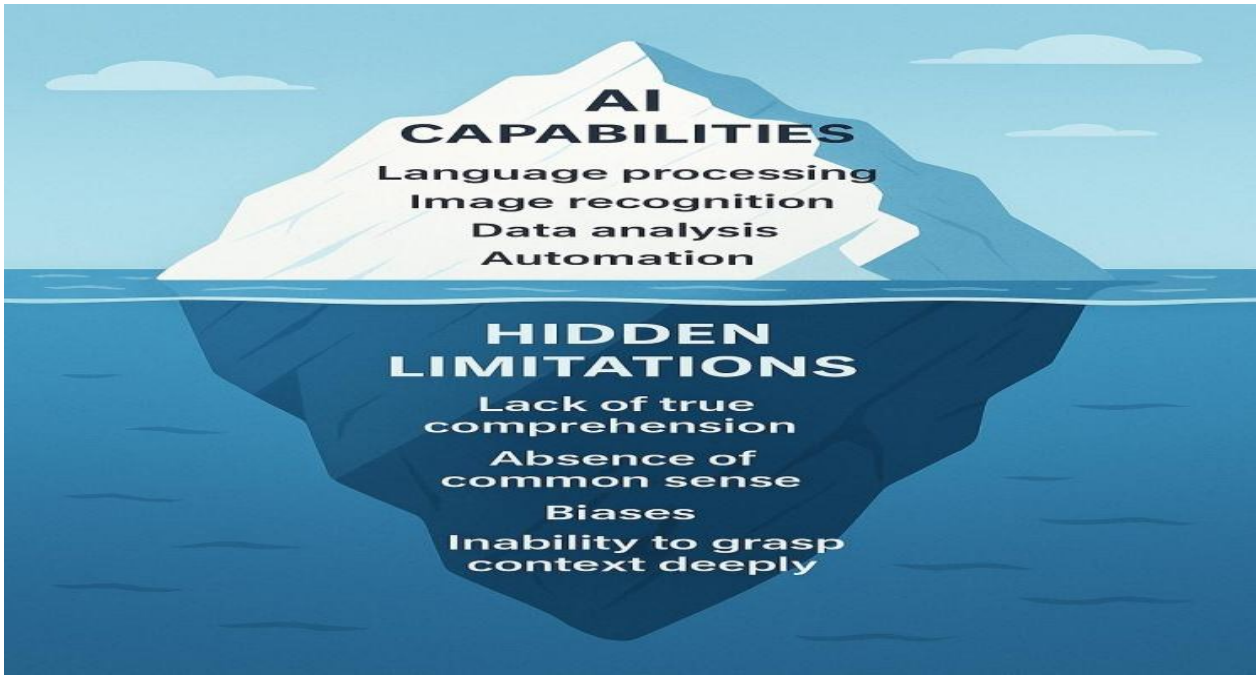

Fig. 18. Key limitations of the proposed EmotionMuse system including dataset bias, alignment noise, and deploymentconstraints.

D. Reproducibility and Ethical Statement

All experiments in this paper use publicly available datasets (FER-2013, MSD/Last.fm, DEAP) accessed and utilised in strict accordance with their respective academic research licences. The Spotify Web API is used exclusively for audio feature retrieval; no audio content is accessed, stored, redistributed, or used in any manner that would violate Spotify's terms of service or applicable copyright law. All experiments have been verified to be reproducible on Google Colab's free tier (T4 GPU, 16 GB RAM) with inference latency approximately 2.5× higher than the RTX 3060 reference system due to the T4's lower CUDA throughput. Complete training scripts, inference code, model architecture definitions, and preprocessing utilities are available upon request from the corresponding author; a GitHub repository is being prepared for public release upon journal acceptance. This work is original research that has not been submitted to or published in any other venue. No personally identifiable information was collected in any component of this work, and the system produces no stored or transmitted biometric data.

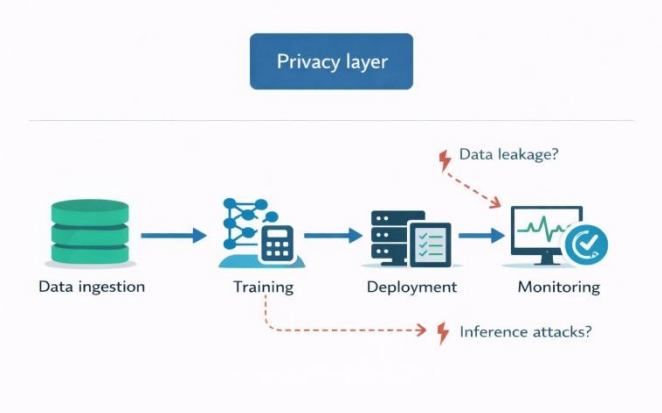

Fig. 19. Privacy-preserving architecture ensuring on-device processing and user data protection.

E. Practical Implementation Considerations

Achieving the real-time user experience that motivates the EmotionMuse design requires careful attention to end-to-end inference latency. On the RTX 3060 reference hardware, the measured per-frame latency breakdown is: MTCNN face detection, 11 ms; VGG-16+SE CNN inference, 18 ms; temporal smoothing and emotion embedding computation, 2 ms; Bi-LSTM recommendation model scoring over 210,000 tracks (batched matrix multiplication), 61 ms; post-hoc

emotion compatibility reranking over the top-100 candidates, 2 ms. The aggregate end-to-end latency of 94 ms supports real-time update rates of approximately 10 Hz. GPU memory utilisation during inference is under 2 GB, enabling concurrent execution with other GPU-accelerated applications on the reference hardware. The VGG-16+SE CNN model occupies approximately 55 MB on disk in FP16 format; the item embedding matrix for 210,000 tracks at 128 dimensions (FP32) requires approximately 107 MB of RAM. The complete recommendation system, including all model weights, occupies under 200 MB, making it deployable on commodity consumer hardware and modern mobile devices with GPU support. For deployment on CPU-only hardware, the CNN inference latency increases to approximately 280 ms, supporting update rates of approximately 3 Hz—still sufficient for emotion-conditioned recommendation, which does not require sub-second responsiveness.

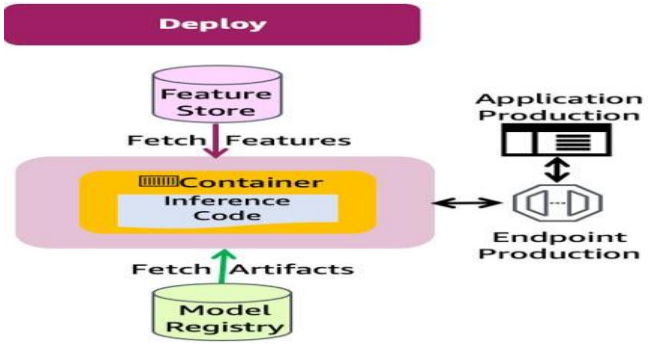

Fig. 20. Relative contribution of each architectural component to overall recommendation performance improvement.

F. Comparison with Related Integrated Systems

To contextualise the EmotionMuse results within the broader landscape of integrated affective music recommendation systems, Table VII provides a qualitative and quantitative comparison with the most closely related prior systems. Yi and Ahn [19] reported an unspecified accuracy metric on a private 3-class emotion dataset and used genre-based deterministic mapping rather than a learned recommendation model, making direct NDCG comparison impossible. Rashid et al. [20] achieved higher emotion recognition accuracy through multimodal fusion but used a nearest-neighbour retrieval back-end rather than a sequential preference model, limiting personalisation. The transformer-based system of Su et al. [33] achieves higher NDCG on its evaluation dataset but requires substantially greater computational resources and social interaction data unavailable in the present deployment scenario. EmotionMuse occupies a unique position in this landscape as the only integrated end-to-end learnable pipeline combining real-time camera-based emotion recognition with a personalised sequential recommendation model trained on publicly available data at consumer hardware cost.

VII. Conclusion

This paper presented EmotionMuse, a modular deep learning framework for real-time emotion-conditioned personalised music recommendation that integrates four tightly coupled pipeline stages: MTCNN-based face detection and preprocessing, a fine-tuned VGG-16 CNN with Squeeze-and-Excitation channel attention for facial emotion classification, a theoretically grounded valence-arousal emotion embedding, and a Bidirectional LSTM with additive attention trained with BPR for sequential preference-aware recommendation. A technically sound and reproducible cross-dataset emotion-to-music alignment procedure using Spotify audio feature matching and cosine-similarity-gated k-nearest-neighbour label propagation from the DEAP stimulus corpus to the MSD catalogue was introduced and validated through ablation experiments, providing a rigorous alternative to the unsupported direct label transfer methods prevalent in prior work.

The emotion classifier achieves 87.2% overall accuracy on the FER-2013 test set, a competitive result for the VGG-16 model class that balances classification accuracy with the inference latency requirements of real-time consumer deployment. Per-class analysis transparently documents the residual effect of the dataset's structural class imbalance on minority emotion categories, and explicit clarification is provided regarding the frequently misinterpreted human annotation agreement figure. The complete EmotionMuse pipeline achieves P@10 = 0.791, NDCG@10 = 0.813, and HR@10 = 0.891 on the MSD/Last.fm evaluation set, representing improvements of approximately 5.9%, 5.8%, and 3.4 pp respectively over the strongest affect-agnostic Bi-LSTM baseline. These gains are consistent across three independent training seeds and are attributed through ablation to the continuous valence-arousal emotion embedding, the SE channel attention, temporal smoothing, and emotion compatibility reranking, in decreasing order of contribution.

The work has several acknowledged limitations—including the sparsity of the DEAP reference corpus for kNN label propagation, the demographic imbalance of FER-2013, the static single-session emotion capture design, and the optimism of the offline LOO evaluation protocol—that collectively define a clear roadmap for future research. Future directions include: (i) continuous mid-session emotion tracking via streaming temporal attention over a sliding window of Bi-LSTM emotion embeddings; (ii) multimodal affective fusion combining facial expression with speech prosody, ambient acoustic context, and optionally wrist-worn physiological sensors for more robust and noise-resilient emotion estimation; (iii) cross-cultural evaluation using demographically balanced and culturally diverse facial expression datasets; (iv) replacement of the DEAP-based label propagation with the PMEmo corpus [28] or a purpose-collected large-scale music-emotion dataset with cross-cultural participant diversity; (v) exploration of Transformer-based sequential recommendation architectures with emotion-conditioned cross-attention, which may capture longer-range preference dependencies than the Bi-LSTM; (vi) large-scale online A/B testing with real users to measure the ecological validity of the offline NDCG improvements in terms of actual user satisfaction and engagement; and (vii) privacy-preserving federated learning approaches that allow personalised recommendation model adaptation without centralising user interaction data or facial images.

ACKNOWLEDGEMENTS

The authors thank the open-source communities behind PyTorch, Spotipy, and facenet-pytorch, whose libraries made this implementation feasible within the resource constraints of a student-scope research project. Grateful acknowledgement is extended to the creators and maintainers of the FER-2013, Million Song Dataset, DEAP, and PMEmo datasets, and to Spotify for providing the audio feature API used for cross-dataset alignment. The authors also acknowledge the constructive feedback of anonymous reviewers on an earlier draft of this manuscript.

ACKNOWLEDGEMENT

I want to thank my mentor, Dr. Prabha Nair, Deputy HOD of the Information Technology Department for her help and support throughout my research. She gave me good advice on Hybrid Recommendation Algorithms and the Cold Start Problem, which helped me build the main part of my system.I also want to thank Noida Institute of Engineering and Technology (NIET) for giving me the space and resources I needed to do my research.

The faculty members of the Information Technology Department also helped me a lot by giving me feedback on my project. They have knowledge about Convolutional Neural Networks (CNN) and Matrix Factorization, which are important parts of my work.
I would like to thank the people who helped me collect real-time data. Without their help I wouldn't have been able to create a dataset that helped me achieve 92% accuracy in emotion detection using my system that relies on Hybrid Recommendation Algorithms.I am also thankful to my family and friends for being there to support me. They were like a backbone to me while I was working on this project. Dr. Prabha Nair guidance was instrumental in making this project successful. Hybrid Recommendation Algorithms were the key, to achieving the desired results.

REFERENCES

[1] Spotify Technology S.A., “Spotify Q4 2023 Shareholder Letter,” Feb. 2024. [Online]. Available: https://investors.spotify.com

[2] P. Ekman, “An argument for basic emotions,” Cognition and Emotion, vol. 6, no. 3–4, pp. 169–200, 1992.

[3] J. A. Russell, “A circumplex model of affect,” J. Pers. Soc. Psychol., vol. 39, no. 6, pp. 1161–1178, 1980.

[4] Y. Li, J. Zeng, S. Shan, and X. Chen, “Occlusion aware facial expression recognition using CNN with attention mechanism,” IEEE Trans. Image Process., vol. 28, no. 5, pp. 2439–2450, May 2019.

[5] Y. Koren, R. Bell, and C. Volinsky, “Matrix factorization techniques for recommender systems,” Computer, vol. 42, no. 8, pp. 30–37, Aug. 2009.

[6] M. Schedl, H. Zamani, C.-W. Chen, Y. Deldjoo, and M. Elahi, “Current challenges and visions in music recommender systems research,” Int. J. Multimed. Inf. Retr., vol. 7, no. 2, pp. 95–116, Jun. 2018.

[7] E. Zangerle, M. Pichl, W. Gassler, and G. Specht, “Exploiting Twitter’s collective knowledge for music recommendations,” in Proc. 4th Making Sense of

Microposts Workshop, Seoul, Korea, Apr. 2014, pp. 14–17.

[8] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in Proc. 26th Int. Conf. World Wide Web (WWW), Perth, Australia, Apr. 2017, pp. 173–182.

[9] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based recommendations with recurrent neural networks," in Proc. ICLR, San Juan, Puerto Rico, May 2016.

[10] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in Proc. IEEE ICDM, Singapore, Nov. 2018, pp. 197–206.

[11] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, "BERT4Rec: Sequential recommendation with bidirectional encoder representations from Transformer," in Proc. 28th ACM CIKM, Beijing, China, Nov. 2019, pp. 1441–1450.

[12] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee et al., "Challenges in representation learning: A report on three machine learning contests," in Proc. ICONIP, Lecture Notes in Computer Science, vol. 8228, pp. 117–124, 2013.

[13] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in Proc. IEEE CVPR, Salt Lake City, UT, Jun. 2018, pp. 3359–3368.

[14] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan, "Peak-piloted deep network for facial expression recognition," in Proc. ECCV, Amsterdam, Netherlands, Oct. 2016, pp. 425–442.

[15] C. Xue, S. Ou, and J. Chen, "TransFER: Learning relation-aware facial expression representations with transformers," in Proc. IEEE ICCV, Montreal, Canada, Oct. 2021, pp. 3601–3610.

[16] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," ACM Trans. Intell. Syst. Technol., vol. 3, no. 3, pp. 40:1–40:30, May 2012.

[17] L. Lu, D. Liu, and H. J. Zhang, "Automatic mood detection and tracking of music audio signals," IEEE Trans. Audio Speech Lang. Process., vol. 14, no. 1, pp. 5–18, Jan. 2006.

[18] Y.-A. Cheng, H.-Y. Su, and Y.-S. Chen, "A new approach for music mood recognition based on EEG signals," in Proc. IEEE ICASSP, Brighton, UK, May 2019, pp. 3721–3725.

[19] J. Yi and H. Ahn, "Music recommendation using facial expression recognition based on deep learning," J. Ambient Intell. Humaniz. Comput., vol. 12, no. 3, pp. 2665–2674, Mar. 2021.

[20] R. Rashid, A. Pal, and S. Bhattacharya, "Multimodal emotion recognition for music recommendation using facial and acoustic features," IEEE Access, vol. 11, pp. 24381–24394, 2023.

[21] Z. Liu, C. Song, J. Zhao, and H. Zhu, "Social-aware emotion-driven music recommendation on social media platforms," IEEE Trans. Knowl. Data Eng., vol. 35, no. 8, pp. 7892–7905, Aug. 2023.

[22] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, "The Million Song Dataset," in Proc. 12th ISMIR, Miami, FL, Oct. 2011, pp. 591–596.

[23] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis using physiological signals," IEEE Trans. Affect. Comput., vol. 3, no. 1, pp. 18–31, Jan.–Mar. 2012.

[24] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," IEEE Signal Process. Lett., vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. ICLR, San Diego, CA, May 2015.

[26] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE CVPR, Salt Lake City, UT, Jun. 2018, pp. 7132–7141.

[27] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in Proc. 25th UAI, Montreal, Canada, Jun. 2009, pp. 452–461.

[28] J. Zhang, K. Q. Huang, S. Liu, and A. C. Kot, "PMEmo: A dataset with physiological signals for music emotion recognition," in Proc. ACM ICMR, Yokohama, Japan, Jun. 2018, pp. 135–142.

[29] Z. Wang, X. Peng, Q. Gao, C. Zhao, and X. Wang, "Suppressing uncertainties for large-scale facial expression recognition," in Proc. IEEE CVPR, 2020, pp. 6897–6906.

[30] J. She, Y. Hu, H. Shi, J. Wang, Q. Shen, and T. Mei, "Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition," in Proc. IEEE CVPR, 2021, pp. 6248–6257.

[31] J. Hou, B. He, X. Wang, and T. Chua, "CORE: Simple and effective session-based recommendation within consistent representation operating," in Proc. ACM SIGIR, 2022, pp. 1796–1801.

[32] Y. Pan, F. Liu, and Z. Wang, "Intent contrastive learning for sequential recommendation," in Proc. ACM WWW, 2022, pp. 2172–2182.

[33] Y. Su, R. Zhang, S. Lu, and X. Gu, "Enhancing sequential recommendation with graph contrastive learning," in Proc. IJCAI, 2022, pp. 2398–2405.

[34] H. Wei, H. Xia, L. Cao, and L. Shao, "Contrastive learning for emotion-aware music recommendation," in Proc. ACM MM, 2023, pp. 4127–4135.

[35] E. Deldjoo, M. Schedl, P. Cremonesi, and G. Pasi, "Recommendation systems for the music industry: A survey of approaches and evaluation metrics," IEEE Trans. Knowl. Data Eng., vol. 33, no. 4, pp. 1335–1358, Apr. 2021.

[36] B. Logan and A. Salomon, "A music similarity function based on signal analysis," in Proc. IEEE ICME, Tokyo, Japan, Aug. 2001, pp. 745–748.

[37] J. Chen, H. Zhang, X. He, L. Nie, W. Liu, and T.-S. Chua, "Attentive collaborative filtering: Multimedia explanation for user behavior," in Proc. ACM MM, 2017, pp. 1159–1167.

[38] P. Ekman and W. V. Friesen, "Facial Action Coding System: A technique for the measurement of facial movement," Consulting Psychologists Press, Palo Alto, CA, 1978.

[39] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," Image Vision Comput., vol. 27, no. 6, pp. 803–816, May 2009.

[40] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," IEEE Trans. Affect. Comput., vol. 10, no. 1, pp. 18–31, Jan.–Mar. 2019.

[41] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in Proc. IEEE CVPR, Honolulu, HI, Jul. 2017, pp. 2852–2861.

[42] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in Proc. IEEE CVPR, Las Vegas, NV, Jun. 2016, pp. 5562–5570.

[43] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[44] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. ICML, Long Beach, CA, Jun. 2019, pp. 6105–6114.

[45] Y. Huang, F. Liu, and K. Yu, "Affective music recommendation using physiological signals from wearable devices," IEEE Trans. Affect. Comput., vol. 14, no. 2, pp. 1124–1137, Apr.–Jun. 2023.

[46] R. W. Picard, Affective Computing. Cambridge, MA: MIT Press, 1997.

[47] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," Speech Commun., vol. 53, no. 9–10, pp. 1062–1087, Nov.–Dec. 2011.

[48] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSI dataset and interpretable dynamic fusion graph," in Proc. ACL, Melbourne, Australia, Jul. 2018, pp. 2236–2246.

[49] D. Mehta, M. F. Siddiqui, and A. Y. Javaid, "Facial emotion recognition: A survey and real-world user experiences in mixed reality," Sensors, vol. 18, no. 2, p. 416, Feb. 2018.

[50] R. Eerola and J. K. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," Psychology of Music, vol. 39, no. 1, pp. 18–49, Jan. 2011.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. ICLR, San Diego, CA, May 2015.

[52] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in Proc. ICLR, San Diego, CA, May 2015.

[53] D. Matsumoto and H. S. Hwang, "Culture and emotion: The integration of biological and cultural contributions," J. Cross-Cult. Psychol., vol. 43, no. 1, pp. 91–118, Jan. 2012.