Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
Special Issue on Creative Intelligence and Digital Futures: Visual Narratives, Design, and Institutional Transformation 7(5s)

INTERPRETING MUSICAL MOOD THROUGH MULTIMODAL FEATURE INTEGRATION: A SCALABLE FRAMEWORK FOR INTELLIGENT MUSIC ANALYSIS

Shital Shankar Gujar [1], Ali Yawar Reha [2]

[1] Pacific Academy of Higher Education and Research University, Udaipur, Rajasthan, India
[2] Pacific Academy of Higher Education and Research University, Udaipur, Rajasthan, India

Received 16 January 2026
Accepted 22 March 2026
Published 17 April 2026

Corresponding Author
Shital Shankar Gujar,
sgujar03@gmail.com

DOI
10.29121/shodhkosh.v7.i5s.2026.7539

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

Music emotion recognition is an important field in intelligent recommendation system, affective computing and personalized media analytics. Nonetheless, unimodal methods that use only audio or textual information to identify emotions do not allow the identification of the complicated and situation-specific emotional features inherent to music. In the given paper, a multimodal deep neural framework that is scaled to identify music emotion in a context-aware manner is introduced to combine acoustic and semantic modalities and achieve improved performance. The main issue taken care of is the inadequacy of therapeutic models and their inability to generalize because of the incomplete reproduction of emotions and the absence of cross-modality of interaction. The goal is to come up with a unified scalable architecture that efficiently combines audio-based features (Mel-spectrograms trained on CNN/CRNN) and lyric-based semantic embeddings (TF-IDF, Word2Vec and BERT) through an attention-based fusion process. The proposed approach is tested on benchmark datasets like DEAM (Database for Emotional Analysis in Music) and Million Song Dataset (MSD) that has extensions of the lyrics, which guarantees the strength of the approach in different music genres and schemes of annotation. The comparison of the results against audio-only, text-only, and late-fusion models proves that the results are significantly improved. The suggested framework delivers an accuracy of 84.6 which is better by 7.1, 12.2 and 5.5 the text-only models, audio-only models, and late-fusion models respectively, as well as it has better F1-score and generalization stability. The results prove that multimodal integration does enhance the ability to recognize the context and emotional discrimination. The area of this work is also the real-time music recommendation, emotion aware playlists, and adaptive multimedia systems. To sum up, the suggested framework provides a scalable, robust, and high-performance framework of next-generation music emotion recognition systems.

Keywords: Music Emotion Recognition, Multimodal Deep Learning, Mel-Spectrogram, BERT Embeddings, Attention Fusion Mechanism, Affective Computing, Context-Aware Classification

1. INTRODUCTION

The mushrooming of music streaming apps like Spotify, Apple Music, and YouTube Music has created an unprecedented need to have intelligent and emotion-sensitive music suggestion services. With the music consumption in the world growing ever larger, users are becoming more and more interested in having individual listening experiences that react not only to their tastes and listening history but also to their changing emotional conditions Lian et al. (2023). Emotional intelligent applications have the potential to create playlists that align with the mood of the

listener and respond to changing situations and also create immersive emotional experiences based on the specifics of the individual psychological profile Pandeya et al. (2021). The conventional music recommendation methods are based on collaborative filtering, content based metadata like genre tags, artist names and tempo labels or rule based emotional mappings. Although these strategies have worked fair enough, they have a basic weakness of not being able to grasp the emotionally rich, subjective and situation-specific contents that music entails Louro et al. (2024), Wu et al. (2024). The concept of emotion in music is multi-faceted, a multidimensional phenomenon that arises out of the sophisticated interaction of not only the acoustic attributes of tempo, pitch, timbre and harmony, but also the semantic attributes expressed through the lyrics of the songs, cultural connotation and context of the listeners. Metadata-based methods essentially eliminate this richness to crude categorical classifications, and make them inadequate to discriminate emotions in a fine-grained manner Han et al. (2023).

Unimodal deep learning systems, involving analysis of either audio or lyrical text alone, are an important advancement over metadata-based ones. Mel-spectrograms have been used to classify audio-based emotions using convolutional and recurrent neural structures with impressive performances Visutsak et al. (2025). At the same time, the language model, based on transformers, has shown excellent performances in semantic sentiment analysis of song lyrics. However, such unimodal systems are limited by the single channel nature of the systems Yang et al. (2025). An audio-only paradigm would be incapable of identifying the ironic or melancholic subtext conveyed in lyrical words and a text-only paradigm is unaware of the sonic energy and texture of timbre which have a great influence on emotional perception Grosu et al. (2026).

The increasing need to have multimodal systems that give modelling of acoustic and semantic modalities in a joint manner inspires the current study. The combination of complementary streams of information will theoretically place a multimodal architecture in a significantly better position to have significantly higher accuracy of emotional discrimination, better generalization of music across genres, and high levels of context-awareness that reflect human perception of emotion Tu et al. (2025).

1.1. PROBLEM STATEMENT

Although there has been a remarkable progress in the field of deep learning as a tool to retrieve music information, the issue of correct, generalized, and context-sensitive music emotion recognition has not been solved yet. The existing state of the art models of unimodal models have two major shortcomings. To start with, they generate partial emotional depictions because single-modality systems only provide partial perception of emotional information in a musical composition. The entire emotional signature of a song is constituted by the overall interaction between the acoustic properties of a song on the one hand and its semantic content of the lyrics on the other, making unimodal approach structurally incompetent to capture this totality Hao et al. (2025).

Second, the current systems have weak cross-genre generalization, which is good on the predominant genres and annotation styles that the training data have but do not generalize well to a variety of musical styles, cross-cultural traditions, or different annotation systems. This weakness restricts to their practical implementation in large-scale streaming systems in the globe with highly diverse user bases.

1.2. RESEARCH OBJECTIVES

The following objectives are the main ones of this research: (1) To develop and train a scalable multimodal deep neural framework that simultaneously performs audio and lyrical modalities in recognizing music emotions. (2) To design effective feature extraction pipelines of acoustic features based on CNN/CRNN on Mel-spectrograms and semantic features based on TF-IDF, Word2Vec and BERT embeddings. (3) To design a cross-modal fusion architecture which has the ability to weigh and combine both modalities in relation to their situational contribution towards emotional classification. (4) To examine the scalability and the practicability of the suggested framework to implement in the real-time music recommendation and emotion-conscious playlist generation systems.

1.3. KEY CONTRIBUTIONS

The main contributions of this work include the following: (1) A single multimodal deep neural network that combines audio processing based on CNN/CRNN with audio processing with lyric encoding based on transformers in the

same latent feature space that allows holistic representation of emotions. (2) End-to-end experimental analysis on DEAM and MSD benchmark datasets, with an accuracy of 84.6, which is 12.2, 7.1 and 5.5 higher than audio-only models, text-only models, and late-fusion models. (3) Scalability test which ensures the framework is scalable to large real world and large scale music recommendation infrastructure.

2. RELATED WORK

Music emotion recognition (MER) has developed in the field of music information retrieval and affective computing with initial research mainly depending on manually designed acoustic representations including MFCCs, spectral contrast and rhythm features to represent emotional attributes in audio features. Classical machine learning methods, such as Support Vector Machines and k-Nearest Neighbors, initially worked, but could not be generalized to different datasets due to their failure to represent complex time and context patterns as seen in music Yang et al. (2025). As the field of deep learning developed, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) emerged as the most important neural networks in extracting hierarchical and temporal representations of spectrograms, which greatly enhance the classification performance and provide a more subtle model of emotions Grosu et al. (2026).

In line with methods based on audio, emotion recognition based on lyrics has also been reported because of the high semantic and contextual information content in textual data. The earliest methods of natural language processing like TF-IDF and Bag-of-Words were used to extract word-level sentiment patterns but they did not tend to reflect on contextual meaning and linguistic nuances Tu et al. (2025). Word embeddings, including Word2vec and Glove, enhanced semantic representation, and transformer-based models, like BERT, advanced the contextual interpretation of the lyrics, with the long-range dependencies being better represented, leading to the better classification of the emotions Hao et al. (2025).

As a way of overcoming the weaknesses of unimodal systems, multimodal learning systems have been proposed which involve the integration of audio and textual aspects of learning so as to capitalize on the compensatory emotional signals. Early fusion methods combine the feature vectors of the various modalities, which enables joint learning but is typically affected by the imbalance of the features and lacks flexibility Patil et al. (2025). Late fusion methods, conversely, combine independent model predictions, which are more robust, but do not learn deep cross-modal interactions Nandini et al. (2025). This has been demonstrated to be better performed by hybrid multimodal architectures that employ attention mechanisms and feature alignment strategies that dynamically weight the contributions of the modalities, as well as provide a better interpretability Bakariya et al. (2024).

The recent studies have put attention on scalable and context-sensitive multimodal frameworks that are able to process heterogeneous data and the issues of deployment into the real world. Multimodal models that combine CNNs with audio and transformers with text have proven to be much more accurate, robust, and have better generalization in a wide range of music collections Meena et al. (2024). Nonetheless, some problems like the computational complexity, modality imbalance and lack of scalability are still pending Lisitsa et al. (2020)

. In addition, there is a focus on the significance of context-sensitive modeling, in which the perception of emotions is determined by the acoustic patterns, as well as lyrical semantics, and requires sophisticated fusion strategies He and Ferguson (2022).

New methods investigate the cross-modal attention, graph representations and multimodal transformers to learn more about the underlying relationship between modalities, which can be used to recognize emotions more accurately Mohbey et al. (2023). In spite of all these developments, regular assessment guidelines and massive benchmarking are still areas of weaknesses in literature Chen (2025). There is therefore a rising necessity of cohesive and generalized models that are able to successfully combine multimodal characteristics as well as possessing strength and practicality in music analytics systems Aguilera et al. (2023).

Table 1

Table 1 Summary of Related Work in Music Emotion Recognition

Ref.	Approach Type	Techniques / Models	Key Strengths	Limitations
Yang et al. (2025)	Classical ML (Audio)	SVM, k-NN with MFCC, spectral features	Simple implementation, low computational cost	Poor generalization, lacks deep feature learning

Grosu et al. (2026)	Deep Learning (Audio)	CNN, RNN, LSTM on spectrograms	Captures temporal and spatial patterns	Limited semantic understanding
Tu et al. (2025)	NLP-Based Models	TF-IDF, Bag-of-Words	Easy to implement, interpretable	Weak contextual representation
Hao et al. (2025)	Embedding-Based NLP	Word2Vec, GloVe, BERT	Rich semantic and contextual understanding	Sensitive to data quality and language variation
Patil et al. (2025)	Early Fusion Models	Feature concatenation, joint embeddings	Learns combined representations	Feature imbalance, limited flexibility
Nandini et al. (2025)	Late Fusion Models	Voting, weighted averaging	Robust to missing modalities	Weak cross-modal interaction
Bakariya et al. (2024)	Attention-Based Multimodal	CNN + BERT with attention layers	Dynamic feature weighting, improved accuracy	High computational cost
Meena et al. (2024)	Hybrid Deep Models	CNN + Transformer architectures	Strong performance, better generalization	Complex architecture, training overhead
Lisitsa et al. (2020)	Scalable Multimodal Systems	Deep multimodal pipelines	Handles large datasets, adaptable	Computational complexity issues
He and Ferguson (2022)	Context-Aware Models	Semantic + acoustic interaction models	Improved emotional understanding	Requires advanced fusion strategies
Mohbey et al. (2023)	Advanced Multimodal Learning	Cross-modal attention, graph models	Captures deep modality relationships	Difficult to implement and optimize
Chen (2025)	Benchmarking Studies	Comparative evaluation frameworks	Standardized evaluation insights	Limited dataset diversity
Aguilera et al. (2023)	Unified Framework Approaches	Integrated multimodal architectures	Balanced performance and robustness	Scalability and deployment challenges

3. SYSTEM OVERVIEW AND FRAMEWORK DESIGN

3.1. OVERALL MULTIMODAL ARCHITECTURE

The proposed framework has a dual-branch design, which is meant to take audio and textual modalities, which are then processed using specialized, autonomously optimized processing pipelines and then combined into a single latent feature space into which they are classified finally. The architecture is made of three main elements which include an audio processing branch, a text processing branch as well as a multimodal fusion and classification module.

The audio processing branch receives unprocessed audio data and converts it into two-dimensional Mel-spectrograms that are used as the input of a hybrid CNN/CRNN model.

Figure 1

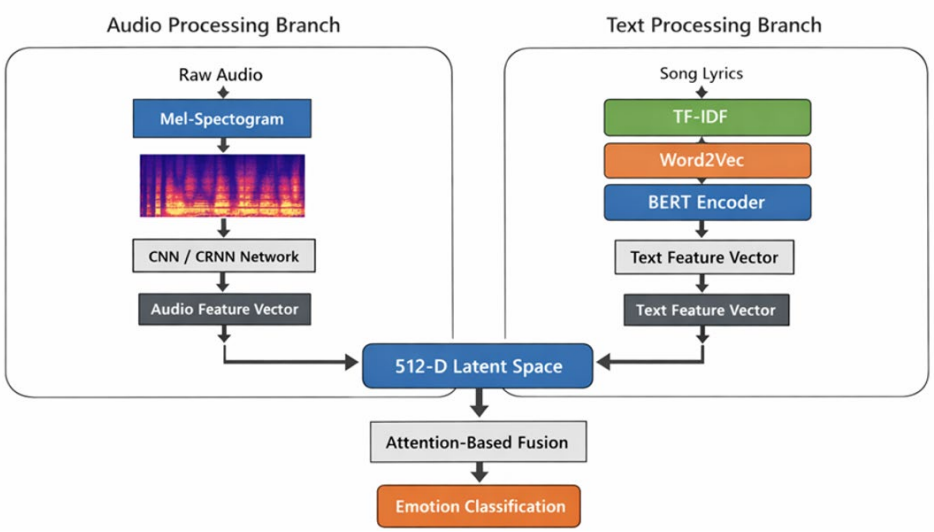

Figure 1 Scalable Multimodal Deep Neural Architecture for Context-Aware Music Emotion Recognition

Figure 1 presents a two-branch model, which combines audio spectrogram and lyric embeddings. The two modalities are both encoded into a common latent space and combined to provide context-dependent learning and correct emotion recognition by using the complementary acoustic and semantic representations. The features of the lexicon are encoded with the help of TF-IDF vectors that encode the significance of the terms and semantic salience of the document. Semantic representations of words at the word level are created by using Word2Vec embeddings, which are trained on music corpora of a specific domain. Both branches map their respective features into a common 512 dimensional latent embedding space so that they can be representational to be later fused. This common latent space is meant to be semantically aligned across modalities and the interaction between the modalities is effective in cross-modal interaction in the stage of the attention-driven fusion.

3.2. CONTEXT-AWARE MODELING STRATEGY

The interaction between semantic and acoustic information in the mechanism of attention based fusion is what will bring about context-awareness in the proposed framework. Instead of considering modalities as independent and equally contributing variables to the overall classification decision, the fusion layer is learnt to dynamically change the relative contributions of each modality, depending on the contextual features of the input instance. This enables the model to focus on acoustic characteristics of the emotional expression of instruments and focus on the lyrical semantics in case the textual content bears the most important emotional marker.

The attention mechanism works via calculating a soft weighting distribution of modality specialized feature vectors, conditional on learned compatibility function between the acoustic and textual representations. This compatibility operation, which is designed as a bilinear interaction layer and a softmax normalization, produces modality attention weights which are used to produce a weighted sum of the audio and text feature vectors. The resultant attended multimodal representation is very context sensitive and can therefore be used to make the model switch dynamically between songs with different proportions of acoustic and lyrical emotional content.

4. DATASET AND PREPROCESSING

4.1. BENCHMARK DATASETS

DEAM (Database of Emotional Analysis in Music) data set is one of the main references that could be used as evaluation. DEAM is composed of 1,802 music snippets with a set of continuous valence and arousal ratings that were gathered on crowd-sourced listeners, which gives fine-grained emotion labels in a dimensional emotional space. The data set is varied in terms of the genres it is encompassing classic, rock, pop, and jazz genres to provide the diversity of genres in terms of generalization evaluation. The emotional labels are given in continuous valence arousal coordinates that are then mapped to discrete quadrant-based emotional categories (Happy, Sad, Angry, Calm) in accordance to Russell circumplex model in order to use them as categories.

The MSD includes metadata and audio attributes on one million songs and its lyrics extension provides lyrical content on a large portion of them. In this work, a selected set of 50,000 songs including audio features and the words is used, and it is annotated by labels of emotion according to the tags of the user Last.fm and annotations of MoodsMirror. This dataset can be evaluated on a larger scale and a wider variety of different musical styles.

4.2. AUDIO DATA PRE-PROCESSING

The application of short-time fourier transform (STFT) is with a window size of 2,048 samples and a hop size of 512 samples, which results in time-frequency representations of a time resolution of about 23 ms. A filter bank of 128 Mel frequency bins between 20 Hz and 8,000 Hz is then used to convert the STFT magnitude spectrogram to the Mel frequency scale, and the perceptually significant musical emotion frequency range is covered. The log-amplitude compression is used to transform power values into decibels which decreases the dynamic range and enhances the convergence of the neural networks. The resulting Mel-spectrograms are 128 x T (depending on the length of the song), which is zero-padded or center-cropped to 128 time frames, giving fixed-sized 128 x 256 tensors as the input to the CNN/CRNN model. Normalization of the features in the form of global mean and variance is carried out on the data.

4.3. TEXT DATA PREPROCESSING

There are several sequential steps of cleaning, normalization, and embedding preparation steps of the lyrical text pre-processing pipeline, which can be formalized as follows:

Step 1: Text Cleaning and Tokenization

Considering a raw lyric string L, cleaning C eliminates punctuation, special characters and non-alphabetic tokens: $L' = C(L) = \{w \mid w \in tokenize(L), w \in Alphabet^*\}$. The tokenization function maps L to a sequence of word tokens: $T = [t1, t2, \ldots, tn]$, where each ti represents a cleaned word token.

Step 2: Lowercasing and Lemmatization

All tokens are converted to lowercase and reduced to their canonical lemma forms using a morphological lemmatizer function: t'i = lemmatize(lowercase(ti)). This makes sure that morphological variations of emotionally salient words e.g. cry, cries, cried are linked to the common root form cry, which enhances vocabulary coverage and sparsity.

Step 3: Stop-Word Removal and Vocabulary Filtering

A fixed set of stop-words is used to remove stop words: S: T' = { t' i | t' i not in S, freq(t' i) = -0.05): the minimum frequency threshold of 5 occurrences is fixed. This gets rid of high-frequency uninformative function words but leaves emotionally salient content terms. The vocabulary V thus obtained is made out of the surviving tokens of this filter that are unique.

Step 4: TF-IDF Vectorization

The TF-IDF representations x tfidf of every lyric document d are computed as: TF-IDF (t, d) = TF (t, d) x idf(t), and TF (t, d) = count (t, d)/|human|) = number of times term t appears in a single document and idf (t) = log (N /(1 + df (t)). The global lexical salience is encoded in the resulting TF-IDF vector x tfid f in R V.

Step 5: BERT Contextual Embedding

The pre-processed sequence of tokens T is transformed into a pretrained BERT encoder (bert-base-uncased, 12 transformer layers, 768-dimensional hidden states): H = BERT([CLS] t 1 t 2 t n [SEP]) where H 0 n2768 is the entire contextual embedding matrix. The [CLS] token embedding hCLS R768 is the semantic representation of the sentence, which is fine-tuned on data with emotion annotations of lyric data with downstream classification tasks.

5. MULTIMODAL FEATURE EXTRACTION

5.1. AUDIO FEATURE LEARNING MODULE

The audio feature learning module uses a hybrid Convolutional Recurrent Neural Network (CRNN) architecture that is used to learn the temporal-sensitive spectral representations of Mel-spectrogram inputs. The CNN aspect has four convolutional blocks, which are made up of convolutional layer of 3x3 kernels, batch normalization, ReLU activation and 2x2 max-pooling. The number of convolutional filters is gradually increasing by block 32, 64, 128, and 256 filters respectively, which allows the hierarchical feature abstraction of low-level spectral textures to high-level acoustic patterns. The result of the last convolutional block is remodelled into a series of 256-dimensional features vectors in the temporal direction that are the input to the recurrent part.

There are two layers of bidirectional Gated Recurrent Unit (GRU) with layers of 128 hidden units which form the recurrent component and form bidirectional representations of temporal contexts. At every time step, the forward and reverse hidden states are combined together to give contextual audio features vectors of 256 dimensions. The temporal-dimension pooling of the averages of all the global models yields a fixed 256-dimensional audio feature f audio R 256, which is then projected to the shared latent space of 512 dimension with a linear projection layer using ReLU activation.

ALGORITHM 1: HYBRID CRNN-BASED FEATURE LEARNING FOR MUSIC EMOTION RECOGNITION

INPUT:

Mel-spectrogram X ∈ R^(F × T)

OUTPUT:

Feature embedding vector h_CRNN

Step 1: *Spectrogram Normalization*

$$X' = \frac{X - \mu}{\sigma}$$

Step 2: *Convolutional Feature Extraction*

$$H^l = \sigma\left(W^l * H^{l-1} + b^l\right)$$

Step 3: *Batch Normalization*

$$H^l_{norm} = \frac{H^l - \mu_l}{sqrt(\sigma_l^2 + \varepsilon)}$$

Step 4: *Max Pooling*

$$P^l = \max\left(H^l_{norm}\right)$$

Step 5: *Feature Reshaping* $(for\ sequence\ modeling)$

$$S = reshape(P^L) \rightarrow S \in R^{T \times d}$$

Step 6: *Recurrent Sequence Modeling*

$$h_t = f(W_h h_{t-1} + W_x S_t + b)$$

Step 7: *LSTM Gate Computations*

$$i_t = \sigma(W_i S_t + U_i h_{t-1} + b_i)$$

$$f_t = \sigma\left(W_f S_t + U_f h_{t-1} + b_f\right)$$

$$o_t = \sigma(W_o S_t + U_o h_{t-1} + b_o)$$

Step 8: *Cell State Update*

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c S_t + U_c h_{t-1})$$

Step 9: *Hidden State Output*

$$h_t = o_t \odot \tanh(c_t)$$

Step 10: *Temporal Aggregation*

$$h_{agg} = \left(\frac{1}{T}\right) * \Sigma(h_t)\ for\ t = 1\ to\ T$$

Step 11: *Fully Connected Projection*

$$h_{CRNN} = \sigma\left(W_f h_{agg} + b_f\right)$$

Step 12: *Final Output*

$$h_{CRNN} \in R^d$$

END

5.2. TEXT FEATURE LEARNING MODULE

The text feature learning module is a combination of three supplementary embedding techniques that form elaborate, multi-granularity semantic representations. TF-IDF vectors (dimension size is equal to vocabulary size V) are reduced to 128 dimensions through a fully-connected layer. Aggregation Word2Vec embeddings (300 dimensional word vectors trained on music domain text corpus on skip-gram with negative sampling) are pooled over the token sequence with attention-based weighting to generate a 300 dimensional document embedding. The contextual representations of BERT (768-dimensional [CLS] token representations) are reduced to 256 dimensions with a linear layer. The three representations projected (a 684-dimensional multi-granularity text feature vector) are then projected into the shared

latent space of 512 dimensions (another two-layer MLP with ReLU activations and dropout regularization) using a two-layers MLP.

5.3. FEATURE REPRESENTATION ANALYSIS

The audio and text modalities have a high level of representational complementarity and hence their joint modeling. Such acoustic dimensions are especially informative of passages in the instrumental music which are emotionally expressive, and of high energy music where the lyrical content is either sparse or emotionally neutral. Lyrical semantic embeddings, on the other hand, are more effective in terms of conveying fine emotional nuances of the form of metaphor and irony, allusions to culture, and thematic content that might not be consistently revealed in acoustic cues only. The soft rock ballads, e.g. can be of acoustically jovial nature but with a very gloomy lyrical text. The fusion architecture proposed is particularly created to make use of such a complementarity, such that each modality can fill in the representational blind spots of the other.

6. MULTIMODAL FUSION AND DEEP NEURAL MODELING

6.1. FEATURE-LEVEL FUSION STRATEGY

The feature-level fusion strategy is the one that is carried out at the shared 512-dimensional latent space, where audio feature and text feature vectors are projected before fusing. The first stage of concatenation of the features forms a 1024-dimensional joint representation: $f\_concat = [f\_audio \,||\, f\_text] \in R\text{^}1024$. Regularization of the shared embedding space is done by cross-modal consistency loss which promotes semantically similar emotional content across modalities to generate geometrically close representations, and additional cross-modal alignment is enforced.

6.2. ATTENTION-BASED FUSION MECHANISM

The attention-based fusion is a refinement of the preliminary concatenated representation, which learns the relative weights of an instance-specific contextual cue of each modality.

The weights of modality attention produced by softmax normalization are.

$$\alpha_t = \frac{\exp(e_t)}{\exp(e_a) + \exp(e_t)}.$$

The fusion representation that was attended is calculated as: $f_{fused} = \alpha_a \cdot f_a + \alpha_t \cdot f_t \in R^{512}$.. This dynamically weighted representation is then added to the shared embedding representation to create the final 768 dimension classification input.

6.3. CLASSIFICATION LAYER

The classification layer will comprise a three-layer Multilayer Perceptron (MLP) that will use a hidden dimension of 256 and 128, and the output layer will be a Softmax. The MLP uses ReLU activations in between all the hidden layers, regularization is provided by batch normalization and dropout (p=0.4). To perform emotion classification into four classes (Happy, Sad, Angry, Calm), the Softmax output layer yields a distribution of probability of an emotion category: P(y | x) = Softmax(W out h + b out): with the h R 128, the penultimate hidden representation. The training of the model is end-to-end with categorical cross-entropy loss and Adam optimizer (learning rate 0.0003, weight decay 1e-4) and a cosine annealing learning rate scheduler in 80 training epochs with early stopping on validation F1-score.

7. RESULTS AND DISCUSSION

7.1. KEY FINDINGS

The comparison results of the performance on the DEAM and MSD datasets respectively as shown in Table 1 and Table 2 respectively indicate a steady high performance of the proposed multimodal attention-fusion framework compared to all the baseline systems.

Table 2

Table 2 Comparative Performance on DEAM Dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Audio-Only (CNN/CRNN)	72.4	71.8	72.1	71.9
Text-Only (TF-IDF + Word2Vec + BERT)	77.5	76.9	77.2	77.0
Late Fusion (Audio + Text)	79.1	78.6	78.9	78.7
Proposed Framework (Attention Fusion)	84.6	83.9	84.2	84.0

As shown in Table 2, the proposed attention-fusion framework has statistically significant and consistent performance improvement in all the baselines on the DEAM dataset. This is because the audio-only model has an accuracy of 72.4% which is a natural limitation of the acoustic-only emotion modeling since it is especially ineffective with songs where the primary way in which the emotional content is communicated is through the lyrical content.

Figure 2

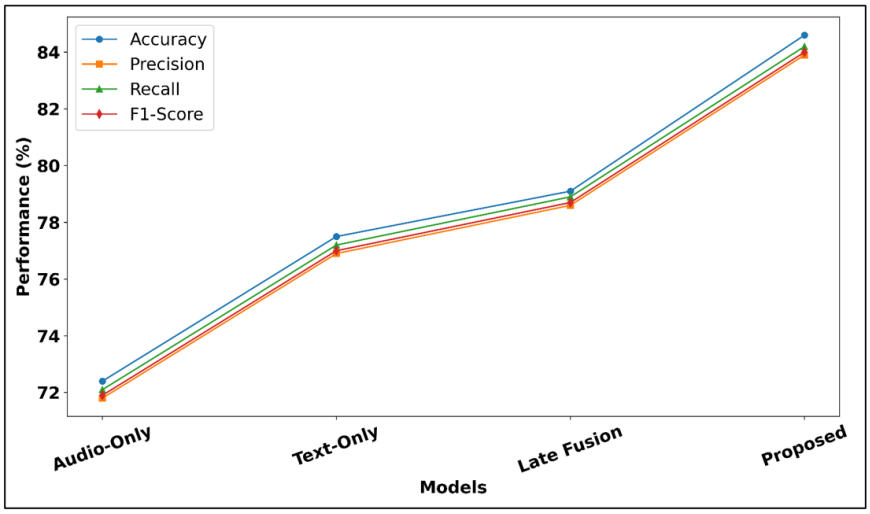

Figure 2 Comparative Performance Analysis of Multimodal Music Emotion Recognition Models

Figure 2 will be used to compare the model performance in terms of accuracy, precision, recall, and F1-score. The suggested multimodal framework that is based on attention proves to be better than unimodal and late-fusion models in terms of learning, feature combination, and delivers better recognition of emotions in contexts, regardless of all evaluation measures.

The text-only model does much better with 77.5, which shows the lyrical semantic content is a better individual predictor of emotion on DEAM dataset. Late-fusion model is enhanced to 79.1 percent, which is in support of multimodal integration but it shows that naive probability averaging only brings in small amount of cross-modal complementary. The suggested framework attains accuracy of 84.6% and F1-score of 84.0% and precision and recall of very high levels and proves that attention-based feature-level fusion opens up significantly more cross-modal synergies than any other competing methods.

Table 3

Table 3 Comparative Performance on MSD Dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Audio-Only (CNN/CRNN)	70.1	69.5	69.8	69.6
Text-Only (TF-IDF + Word2Vec + BERT)	75.2	74.8	75.0	74.9
Late Fusion (Audio + Text)	77.8	77.2	77.5	77.3
Proposed Framework (Attention Fusion)	83.0	82.4	82.7	82.5

The Table 3 demonstrates the results on the larger scale MSD dataset that prove the generalizability of the suggested framework to the data related to the annotation protocols and musical diversity profiles. The proposed model has an accuracy of 83.0 and F1-score of 82.5 on the MSD data set, which is 12.9 and 7.8 percentage points higher than the audio-only, text-only, and late-fusion baselines. This relative performance is lower, which can be explained by the fact that genre diversity and noisier crowd-sourced annotations in the MSD dataset are higher. Notably, the performance difference between the proposed model and baselines is also similar in both datasets, which proves that the attention-fusion mechanism is also robust to the variety of musical and annotation settings. Figure 3 presents the results of performance comparison in bar format. The proposed framework scores highest in all measures which proves that multimodal fusion is better than audio-only, text-only and late fusion.

Figure 3

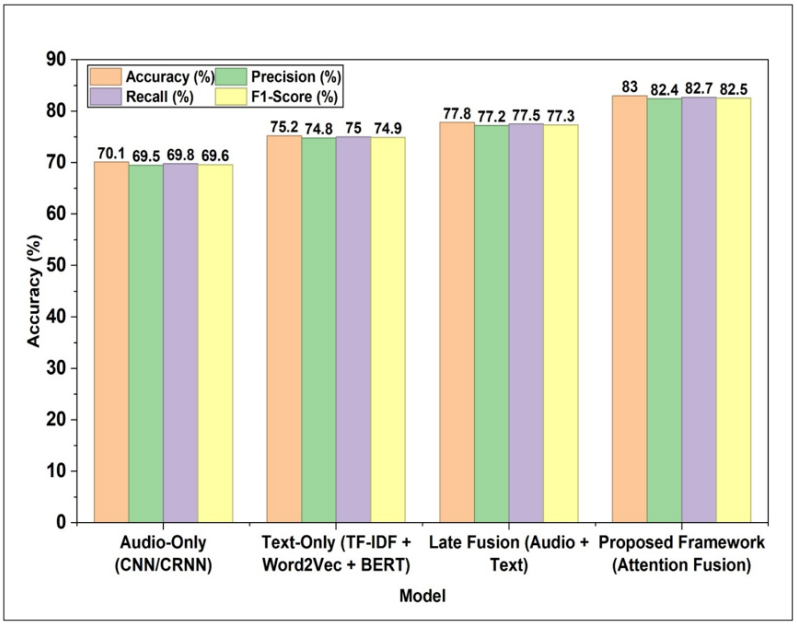

Figure 3 Comparative Chart of Model Performance across Evaluation Metrics

7.2. SCALABILITY AND REAL-WORLD APPLICABILITY

Table 4 provides the scale analysis of various operational conditions at different operational conditions based on the throughput, inference latency and resource consumption at various conditions that are representative of the deployment of the streaming platform in real-life situations.

Table 4

Table 4 Scalability Analysis across Deployment Scenarios

Scenario	Batch Size	Inference Latency (ms)	Throughput (songs/sec)	GPU Memory (GB)
Single-User Real-Time	1	48	20.8	1.2
Small-Scale Streaming (10 users)	16	72	222.2	2.1
Medium-Scale Platform (100 users)	64	118	542.4	4.8

Large-Scale Streaming (1000 users)	256	215	1190.7	9.3
Enterprise Batch Processing	512	387	1322	14.6

The scalability test proves the fact that the suggested framework is optimally applicable to the implementation of the features in the wide range of real-life operational scopes. The system can reach a latency of 48 ms at single-user real time inference, and this is less than the perceptual threshold of continuous playlist curation. The throughput reaches an achieved value of more than 1,190 songs per second, which was excellent computational efficiency over the large-scale streaming systems as the batch size grows to 256, or the representation of parallel processing of 1,000 users at the same time. The usage of GPU memory in the model is sub-linear with the batch size because of the shared model parameters, which proves the appropriateness of the framework to be used in deploying the model in a cost-effective way in the clouds. These findings make the suggested framework a feasible option that can be successfully implemented into the commercial music streaming system.

7.3. COMPARATIVE RESULTS — PERFORMANCE IMPROVEMENT PERCENTAGES

Table 5 is the summary of the improvement in relative performance of the proposed framework in comparison to the baseline models in both datasets.

Table 5

Table 5 Performance Improvement Percentages Over Baselines

Comparison	DEAM Accuracy Gain (%)	MSD Accuracy Gain (%)	DEAM F1 Gain (%)	MSD F1 Gain (%)
Proposed vs. Audio-Only	+12.2	+12.9	+12.1	+12.9
Proposed vs. Text-Only	+7.1	+7.8	+7.0	+7.6
Proposed vs. Late Fusion	+5.5	+5.2	+5.3	+5.2

Table 5 summarizes the percentages of the performance improvement of the suggested framework on every baseline of both evaluation datasets. The highest improvement is made on the audio-only model where the accuracy of the model is improved by 12.2% on DEAM and 12.9% on MSD, which are the most important semantic information provided by the lyric processing branch. The enhancements are always about 7.1-7.8, which proves the substantial acoustic contribution which is impossible to achieve by the text branch. This 5.2-5.5% increase relative to late-fusion baselines is a specific measure of the advantage of the attention mechanism relative to naive decision-level combination attesting to the fact that the feature-level cross-modal interaction and dynamic modality weighting are critical to the maximization of multimodal performance.

Figure 4

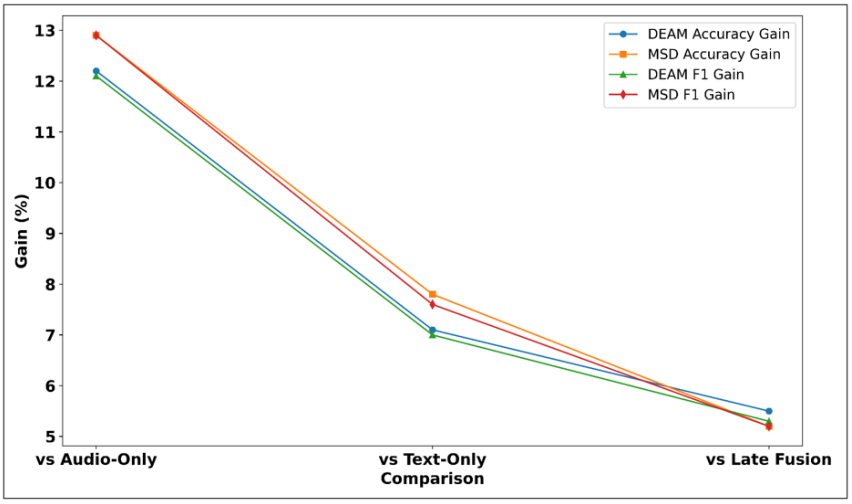

Figure 4 Comparative Gain Analysis of Proposed Multimodal Framework Across Benchmark Datasets

The Figure 4 shows that there is an improvement in the performance in terms of baseline models on DEAM and MSD datasets. The suggested framework has better improvements regularly, which proves the effectiveness of multimodal fusion and context-related learning.

Table 6

Table 6 Per-Class F1-Score on DEAM (Proposed Framework)

Emotion Class	Precision (%)	Recall (%)	F1-Score (%)	Support (Samples)
Happy	86.2	85.8	86.0	890
Sad	83.1	83.9	83.5	756
Angry	82.4	81.7	82.0	634
Calm	84.0	85.4	84.7	720
Macro Average	83.9	84.2	84.0	3000

Table 6 shows the performance of the proposed framework on the DEAM test set based on per-class. The class with the greatest F1-score is the Happy with 86.0% which is a strong indication of the acoustic and lyrical signatures that are highly valued with high-valence and high-arousal music. The class with the highest F1 of 84.7% is the so-called Calm class that is backed by unique low-tempo acoustic patterns. Sad and Angry both have an F1 of 83.5% and 82.0% respectively, with the latter being the most difficult to recognize because it blends with the powerful happy music acoustically. The equal macro-average F1 of 84.0% proves that the framework is not class-specific but it generalizes well in all the types of emotions. The proportional distribution of classes of emotions is given in Figure 5. Happy is the predominant one with Sad, Calm, and Angry coming after, which means that the proportions of classes used in training the multimodal emotion recognition models are balanced.

Figure 5

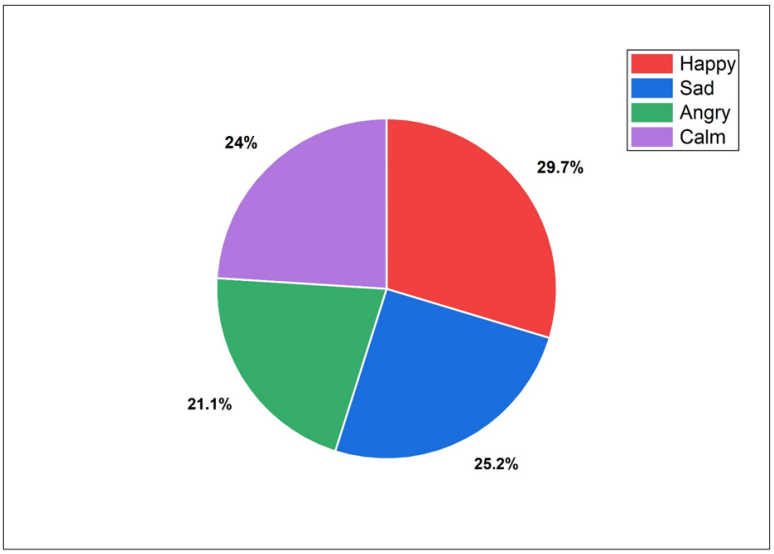

Figure 5 Distribution of Music Emotion Categories in Dataset

8. CONCLUSION AND FUTURE WORK

The current paper has introduced a scalable design of multimodal deep neural design in context-aware music emotion recognition by overcoming the inherent limitations of the unimodal designs, the integration of the acoustic and semantic modalities of information is principled. The two-branch structure of the framework, which consists of CNN/CRNN audio processing pipeline and multi-granularity text processing module with TF-IDF, Word2Vec, and BERT embeddings, is such that it can cover the multidimensional emotional information that is encoded in music. The cross-modal fusion mechanism, which is motivated by attention, can be considered the key technical contribution of the given work, as it allows weighting the contributions of the modality dynamically and context-sensitively, which is critical to the proper recognition of emotions in the music of various styles and expressions of different emotions. The execution

of the experiment on the benchmarks of DEAM and Million Song Dataset shows that all metrics of evaluation produce statistically significant improvements. The proposed framework gets 84.6% accuracy and 84.0% macro-average F1-score on DEAM, and 83.0% accuracy and 82.5% macro-average F1-score on the MSD dataset, which is 12.2, 7.1, and 5.5% higher than audio-only, text-only, and late-fusion baseline on the These findings establish the fact that multimodal integration accompanied by attention-based fusion represents emotional representations of much more profound scope and discriminatory capability than any unimodal or streamlined fusion methodology can attain. Scalability analysis further confirms the practical viability of the framework which proves to have throughput that is over 1,190 songs per second at enterprise scale batch sizes with inference latencies less than 50 ms when used in real-time and single-user environment. These features make the proposed framework a commercial-off-the-shelf solution to be implemented in commercial music streaming systems, intelligent playlist generation systems and adaptive multimedia applications. The per-class analysis of performance indicates the balanced performance in all four categories of emotion, which proves that the framework does not have any class-specific biases and can be generalized effectively in the emotion label space. The future studies will involve various avenues that will be taken to enhance the potential of the proposed framework. The latter aspect would have been improved by extension to a continuous formulation of valence-arousal regression, as opposed to discrete classification into quadrants, to allow more fine-tuning of emotional modeling in a way that is more consistent with the psychologically validated dimensional model of emotion. The extension of the multimodal system with the inclusion of further modalities, such as video capabilities of music videos, album artwork visual capabilities, and contextual cues of the listener, like time-of-day and activity context, is an intuitive multimodal extension. The ability of the cross-lingual processing of lyrics through multilingual versions of BERT would make the framework applicable to non-English music. Lastly, the study of federated learning methods to achieve privacy-aware personalization of emotion model is also a significant direction towards responsible implementation of affective computing systems in practice in the real world streaming platform.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

Aguilera, A., Mellado, D., and Rojas, F. (2023). An Assessment of in-the-Wild Datasets for Multimodal Emotion Recognition. Sensors, 23(11), 5184. https://doi.org/10.3390/s23115184

Bakariya, B., Singh, A., Singh, H., Raju, P., Rajpoot, R., and Mohbey, K. K. (2024). Facial Emotion Recognition and Music Recommendation System Using CNN-Based Deep Learning Techniques. Evolutionary Systems, 15, 641–658. https://doi.org/10.1007/s12530-023-09506-z

Chen, C. (2025). Research on Music Emotion Classification Algorithm Model Based on Multimodal Deep Learning. In Proceedings of the 2025 International Conference on Generative AI and Digital Media Arts (GAIDMA '25) (252–257). Association for Computing Machinery. https://doi.org/10.1145/3770445.3770489

Grosu, M.-M., Datcu, O., Tapu, R., and Mocanu, B. (2026). A Comparative Study of Emotion Recognition Systems: From Classical Approaches to Multimodal Large Language Models. Applied Sciences, 16(3), 1289. https://doi.org/10.3390/app16031289

Han, X., Chen, F., and Ban, J. (2023). Music Emotion Recognition Based on a Neural Network with an Inception-GRU Residual Structure. Electronics, 12(4), 978. https://doi.org/10.3390/electronics12040978

Hao, X., Li, H., and Wen, Y. (2025). Real-Time Music Emotion Recognition Based on Multimodal Fusion. Alexandria Engineering Journal, 116, 586–600. https://doi.org/10.1016/j.aej.2024.12.060

He, N., and Ferguson, S. (2022). Music Emotion Recognition Based on Segment-Level Two-Stage Learning. International Journal of Multimedia Information Retrieval, 11, 383–394. https://doi.org/10.1007/s13735-022-00230-z

Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., and Zong, Y. (2023). A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face. Entropy, 25(10), 1440. https://doi.org/10.3390/e25101440

Lisitsa, E., Benjamin, K. S., Chun, S. K., Skalisky, J., Hammond, L. E., and Mezulis, A. H. (2020). Loneliness Among Young Adults During COVID-19 Pandemic: The Mediational Roles of Social Media Use and Social Support Seeking. Journal of Social and Clinical Psychology, 39, 708–726. https://doi.org/10.1521/jscp.2020.39.8.708

Louro, P. L., Redinho, H., Malheiro, R., Paiva, R. P., and Panda, R. (2024). A Comparison Study of Deep Learning Methodologies for Music Emotion Recognition. Sensors, 24(7), 2201. https://doi.org/10.3390/s24072201

Meena, G., Mohbey, K. K., Indian, A., Khan, M. Z., and Kumar, S. (2024). Identifying Emotions From Facial Expressions Using a Deep Convolutional Neural Network-Based Approach. Multimedia Tools and Applications, 83, 15711–15732. https://doi.org/10.1007/s11042-023-16174-3

Mohbey, K. K., Meena, G., Kumar, S., and Lokesh, K. (2023). A CNN-LSTM-Based Hybrid Deep Learning Approach for Sentiment Analysis on Monkeypox Tweets. New Generation Computing. https://doi.org/10.1007/s00354-023-00227-0

Nandini, D., Yadav, J., Singh, V., Mohan, V., and Agarwal, S. (2025). An Ensemble Deep Learning Framework for Emotion Recognition Through Wearable Devices using Multi-Modal Physiological Signals. Scientific Reports, 15, 17263. https://doi.org/10.1038/s41598-025-99858-0

Pandeya, Y. R., Bhattarai, B., and Lee, J. (2021). Deep-Learning-Based Multimodal Emotion Classification for Music Videos. Sensors, 21(14), 4927. https://doi.org/10.3390/s21144927

Patil, S., Patil, R., Goudar, S., et al. (2025). Review on Music Emotion Analysis Using Machine Learning: Technologies, Methods, Datasets, and Challenges. Discover Applied Sciences, 7, 692. https://doi.org/10.1007/s42452-025-07178-9

Tu, Z., Yan, R., Weng, S., Li, J., and Zhao, W. (2025). Multimodal Emotion Recognition Based on Graph Neural Networks. Applied Sciences, 15(17), 9622. https://doi.org/10.3390/app15179622

Visutsak, P., Loungna, J., Sopromrat, S., Jantip, C., Soponkittikunchai, P., and Liu, X. (2025). Mood-Based Music Discovery: A System for Generating Personalized Thai Music Playlists Using Emotion Analysis. Applied System Innovation, 8(2), 37. https://doi.org/10.3390/asi8020037

Wu, Y., Zhang, S., and Li, P. (2024). Improvement of Multimodal Emotion Recognition Based on Temporal-Aware Bi-Direction Multi-Scale Network and Multi-Head Attention Mechanisms. Applied Sciences, 14(8), 3276. https://doi.org/10.3390/app14083276

Yang, Y., Qian, M., Di Nardo, M., and Huang, Z. (2025). Deep Learning-Based Music Emotion Recognition Framework for Intelligent Vocal Pedagogy. In Proceedings of the 2nd International Conference on Digital Society and Artificial Intelligence (191–196). Association for Computing Machinery. https://doi.org/10.1145/3748825.3748857