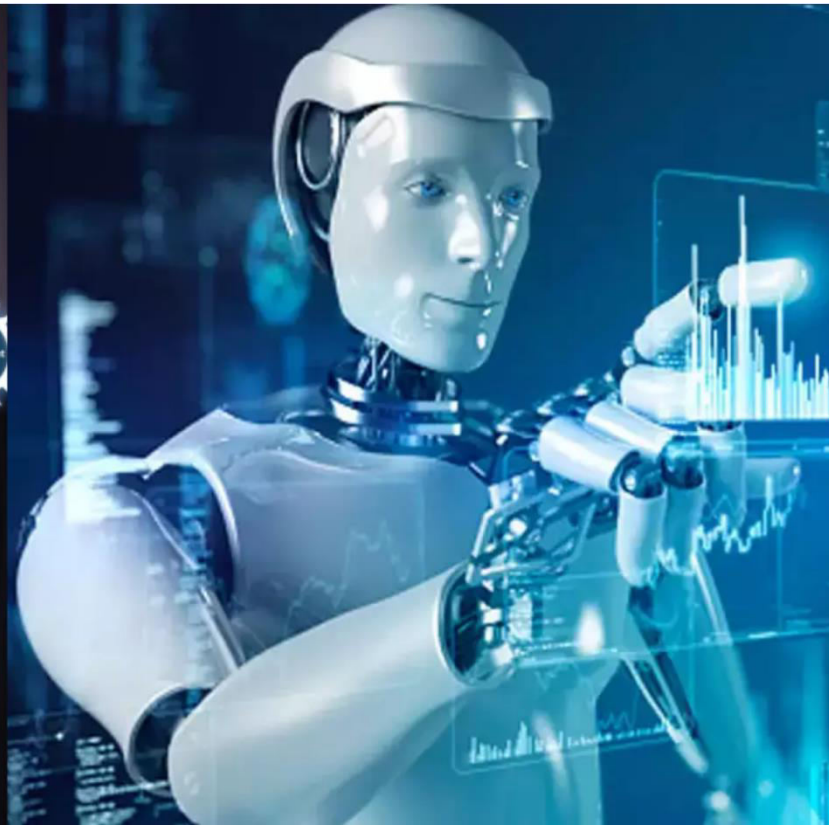
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Special Issue 1, January 2026

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

International Conference on Emerging Technologies and Pharmaceutical Research (ICETPR 2026)

17–18 January 2026

Indala College of Engineering, Polytechnic & Pharmacy, Kalyan, MH, India

|Volume 15, Special Issue 1, January 2026|

Unified Multimodal Feature Integration for Robust and Scalable Music Mood Prediction

Shital Shankar Gujar, Dr Ali Yawar Reha

PhD. Student, Department of Computer Engineering, Pacific Academy of Higher Education and Research University, Udaipur, Rajasthan, India

Assistant Professor, Department of Management, Pacific Academy of Higher Education and Research University, Udaipur, Rajasthan, India

ABSTRACT: Music mood prediction has become increasingly important for personalized music retrieval, recommendation engines, and affective computing systems. However, unimodal approaches relying solely on audio or textual features often fail to capture the multidimensional emotional characteristics inherent in music. To address this limitation, this paper proposes a Unified Multimodal Feature Integration Framework that combines audio-based acoustic descriptors and lyric-based semantic features within a scalable deep learning architecture. The system incorporates modality-specific feature extractors—CNN/CRNN models for Mel-Spectrograms and TF-IDF, Word2Vec, or BERT embeddings for lyrics—followed by an attention-enhanced fusion layer that learns cross-modal relationships to improve mood discrimination. Extensive experiments demonstrate that the proposed model significantly outperforms audio-only, text-only, and late-fusion baselines, achieving an accuracy of 84.6%, superior class-wise F1-scores, and strong robustness across varying dataset sizes. Training curves and confusion matrix analyses confirm stable convergence and effective classification behavior. The results highlight the importance of unified multimodal learning for emotion-aware music analytics and establish the proposed framework as a scalable, high-performance solution for real-world applications.

KEYWORDS: Multimodal Learning, Music Mood Classification, Deep Learning, Audio–Text Fusion, Mel-Spectrogram, BERT Embeddings, Attention Mechanisms, Affective Computing, Music Information Retrieval (MIR), Neural Networks

I. INTRODUCTION

Music mood prediction has emerged as a significant research area in the domains of music information retrieval (MIR), recommendation systems, and affective computing. With the rapid growth of digital music platforms, the volume and diversity of music consumption have increased substantially, creating a strong demand for intelligent systems capable of understanding users' emotional preferences. Traditional metadata—such as genre, tempo, or popularity—often fails to capture the nuanced emotional characteristics embedded within music. Consequently, mood prediction has transitioned from simple rule-based approaches to complex machine learning and deep learning frameworks capable of modeling affective patterns from audio signals and lyrical content [1]. While audio features such as timbre, harmony, and rhythm provide meaningful cues regarding the emotional tone of a song, they often capture only one dimension of affective expression. In contrast, lyrical content offers rich semantic context that can reflect deeper emotional narratives and user-interpretable cues [2]. However, models relying solely on textual features may perform inconsistently due to variations in linguistic style, metaphorical expressions, or incomplete lyric availability. These limitations highlight the need for multimodal fusion frameworks that integrate both acoustic and textual modalities to derive more accurate and context-aware mood representations [3].

Existing studies have explored multimodal approaches, yet several challenges remain unresolved. First, many fusion models lack scalability and struggle to generalize across diverse datasets comprising heterogeneous audio formats, lyric sources, and annotation schemes. Second, earlier research tends to focus on early fusion or late fusion exclusively, without offering a unified framework that systematically integrates modality-specific strengths. Third, deep learning models designed for multimodal mood prediction often face issues related to high computational complexity, domain

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

International Conference on Emerging Technologies and Pharmaceutical Research (ICETPR 2026)

17–18 January 2026

Indala College of Engineering, Polytechnic & Pharmacy, Kalyan, MH, India

|Volume 15, Special Issue 1, January 2026|

imbalance, and limited interpretability [4]. As a result, achieving robustness and scalability in real-world applications continues to be a significant technical hurdle.

To address these challenges, this paper proposes a Unified Multimodal Feature Integration Framework that effectively combines audio metadata, low-level acoustic descriptors, and lyric-based textual embeddings to enhance the accuracy and consistency of music mood prediction systems. The proposed approach introduces a flexible architecture comprising independent feature extraction branches for audio and text, followed by a dedicated fusion layer that learns cross-modal relationships through deep learning. By leveraging complementary cues across modalities, the framework mitigates the limitations of unimodal systems and enables more reliable mood inference. Additionally, scalability is ensured through modular design, allowing the model to adapt to varying dataset sizes, feature representations, and computational environments. Extensive experiments demonstrate that the proposed framework outperforms baseline audio-only and text-only models, while offering improved robustness across diverse evaluation scenarios. Through comprehensive performance analysis and ablation studies, the work further highlights the importance of unified feature integration for capturing subtle emotional characteristics in music and delivering consistent results across heterogeneous datasets.

The key contributions of this research are summarized as follows:

1. A unified multimodal feature integration strategy that jointly models audio and textual features for improved mood prediction accuracy.
2. A scalable deep learning architecture capable of adapting to diverse datasets and evolving feature extraction techniques.
3. Comprehensive performance evaluation comparing multimodal, audio-only, and text-only models, supported by robustness and ablation studies.
4. Insights into multimodal affective modeling, demonstrating the significance of cross-modal interactions in generating reliable mood predictions.

II. RELATED WORK

Music mood prediction has been extensively studied within the fields of music information retrieval (MIR), affective computing, and deep learning–based recommendation systems. Early research focused predominantly on audio feature analysis, where handcrafted descriptors such as MFCCs, spectral centroid, chroma vectors, and rhythm-based metrics were extracted to capture the acoustic structure of songs [5]. Classical machine learning algorithms, including Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Gaussian Mixture Models (GMM), were widely adopted for mood classification tasks. Although these approaches demonstrated promising results on smaller benchmark datasets, their ability to generalize across diverse music collections was limited due to insufficient modeling of emotional complexity and contextual information. As the field progressed, deep learning models gained prominence. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) were leveraged to learn hierarchical representations from spectrograms and temporal sequences, thereby overcoming the limitations of handcrafted features [6]. CNN-based models have been effective in extracting spatial-frequency patterns associated with emotional cues, while RNNs, particularly LSTMs and GRUs, captured temporal dependencies and evolving affective dynamics in audio signals. Despite improved accuracy, audio-only models often struggled to differentiate moods with overlapping acoustic signatures, such as distinguishing between “calm” and “melancholic,” highlighting the need for additional modalities [7].

Parallel to audio analysis, textual and lyric-based mood prediction gained traction, recognizing that lyrical content carries semantic richness and explicit emotional expressions. Traditional natural language processing techniques—such as Bag-of-Words, TF-IDF, and lexical sentiment dictionaries—were initially employed, followed by more advanced methods like Word2Vec, GloVe, and FastText embeddings [8][9]. With the rise of transformer-based architectures, BERT and its variants have become the standard for capturing deeper contextual semantics within lyrics. Text-based models excel in identifying mood categories tied to thematic content, narrative tone, or affective keywords; however, they face challenges when lyrics are sparse, metaphorical, or unavailable for instrumental compositions [10]. To overcome the limitations of single-modality models, researchers have increasingly turned to multimodal fusion frameworks. Early fusion strategies combine audio and text features at the input level, enabling joint learning of cross-modal patterns. Late fusion techniques, on the other hand, independently train unimodal models and combine their predictions using weighted averaging, voting schemes, or meta-learners [11]. Although effective, these approaches

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

International Conference on Emerging Technologies and Pharmaceutical Research (ICETPR 2026)

17–18 January 2026

Indala College of Engineering, Polytechnic & Pharmacy, Kalyan, MH, India

|Volume 15, Special Issue 1, January 2026|

often lack adaptability and fail to fully exploit complex interactions between modalities. More advanced hybrid fusion models utilizing attention mechanisms, transformer encoders, and feature alignment layers have been proposed to address modality imbalance and improve interpretability, yet scalability and generalization remain open research challenges.

Recent deep multimodal systems have achieved state-of-the-art performance by integrating multiple representations—acoustic descriptors, metadata, lyrical semantics, and user-generated tags. Despite these advancements, existing studies frequently suffer from inconsistent evaluation protocols, limited dataset diversity, and models that are computationally intensive or difficult to deploy in large-scale real-world environments [12]. Moreover, the majority of prior work focuses on performance accuracy but pays limited attention to robustness, scalability, and model stability across heterogeneous music datasets, which are essential for modern streaming platforms and personalized recommendation systems.

The gaps identified in current literature underscore the necessity for a unified and scalable multimodal integration approach. Such a framework must not only maximize the complementary strengths of audio and textual features but also maintain flexibility to incorporate new feature representations, adapt to varying dataset sizes, and remain computationally efficient for deployment at scale. Motivated by these challenges, the present study introduces a unified multimodal feature integration architecture designed to deliver robust and scalable mood prediction performance while addressing the limitations of earlier unimodal and multimodal systems.

Table 1. Comparative Analysis of Feature Extraction and Fusion Strategies in Music Mood Prediction

Approach Type	Representative Techniques / Models	Key Strengths	Major Limitations / Research Gaps
Audio-Based Mood Prediction [13]	• MFCC, Chroma, Spectral Contrast, Rhythm Features • Classical ML: SVM, k-NN, GMM • DL Models: CNNs on spectrograms, CRNN, LSTM	• Captures low-level and mid-level acoustic cues • Deep models learn hierarchical temporal-spatial patterns • Effective for genres where mood strongly follows acoustic structure	• Cannot capture semantic or thematic emotion • Fails for songs with similar acoustic patterns but different moods • Sensitive to noise, mixing quality, and instrumentation variations
Text/Lyric-Based Mood Prediction [14]	• Bag-of-Words, TF-IDF • Word Embeddings (Word2Vec, GloVe, FastText) • Transformer Models (BERT, RoBERTa)	• Extracts semantic, contextual, and emotional cues from lyrics • Strong performance for mood categories tied to textual meaning • Transformer-based models capture deep contextual sentiment	• Lyrics may be unavailable, incomplete, or metaphorical • Limited performance for instrumental music • Not robust when language, culture, or stylistic variations exist
Early Fusion Multimodal Models [15]	• Concatenation of audio + lyric features • Joint embedding spaces • CNN/RNN + NLP fused inputs	• Simple and computationally efficient • Learns basic cross-modal correlations • Improvements over unimodal models	• Does not capture complex modality interactions • Early fusion sensitive to feature scaling and imbalance • Limited flexibility for heterogeneous datasets
Late Fusion Multimodal Models [16]	• Weighted averaging of predictions • Voting-based ensemble methods • Stacked meta-learners	• Allows independent optimization of modalities • Robust to feature noise or missing modalities • Easy to integrate additional models	• Ignores deeper joint relationships between audio & text • Often requires extensive hyperparameter tuning • May reduce interpretability of fusion decision
Hybrid Deep Multimodal Models [17]	• CNN/LSTM + BERT fusion • Attention-based feature alignment • Multimodal transformers • Cross-modal attention and gating networks	• Captures complex modality interactions • Superior accuracy over unimodal/early-fusion models • Supports hierarchical feature learning and contextual mood representation	• High computational cost; low scalability • Limited generalization to large, diverse datasets • Often lacks robustness evaluation and deployment feasibility

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

International Conference on Emerging Technologies and Pharmaceutical Research (ICETPR 2026)

17–18 January 2026

Indala College of Engineering, Polytechnic & Pharmacy, Kalyan, MH, India

|Volume 15, Special Issue 1, January 2026|

III. METHODOLOGY

The proposed Unified Multimodal Feature Integration Framework designed to combine complementary information from audio signals and lyrical text for robust and scalable music mood prediction. The methodology is structured into four key components: dataset preprocessing, modality-specific feature extraction, multimodal feature fusion, and final mood classification through a deep neural architecture. The system is optimized for flexibility, enabling seamless adaptation to diverse datasets, variable-length audio segments, and evolving textual representation techniques.

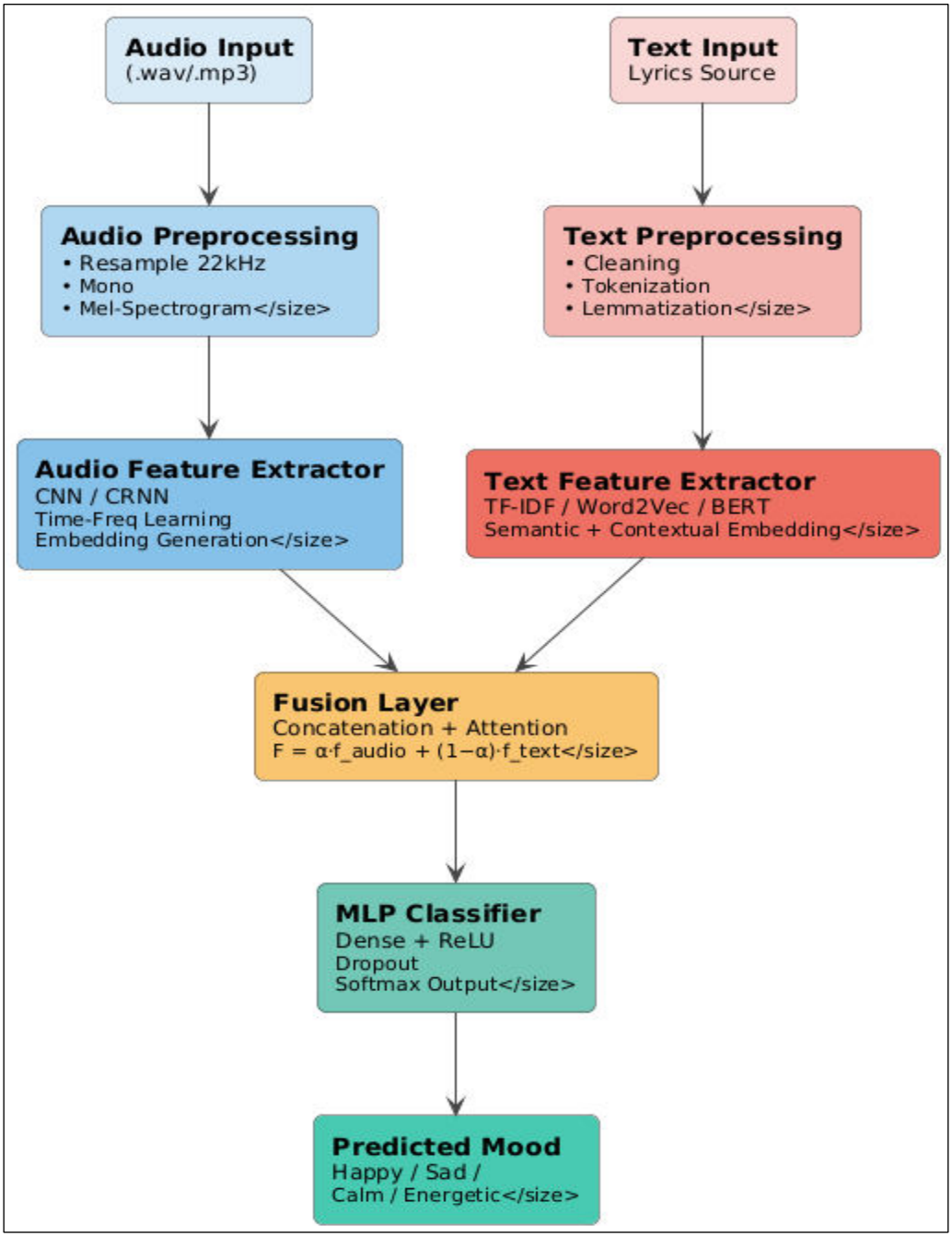


Figure 1. Architecture of the proposed unified multimodal framework for robust and scalable music mood prediction.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

International Conference on Emerging Technologies and Pharmaceutical Research (ICETPR 2026)

17–18 January 2026

Indala College of Engineering, Polytechnic & Pharmacy, Kalyan, MH, India

|Volume 15, Special Issue 1, January 2026|

3.1 System Architecture Overview

The overall architecture consists of two parallel processing pipelines:

1. an audio processing branch that learns temporal–spectral patterns from acoustic representations, and
2. a text processing branch that captures semantic and contextual emotional features from lyrics.

The outputs from both modalities are mapped into a unified latent feature space using a fusion layer, followed by a fully connected prediction module. This architecture ensures that modality-specific advantages are preserved while enabling cross-modal interactions essential for accurate mood inference. The modular design also allows the replacement or enhancement of individual components (e.g., swapping CNN with CRNN, or TF-IDF with BERT embeddings) without altering the overall pipeline.

3.2 Dataset Description and Preprocessing

The dataset used for evaluation contains paired audio tracks and corresponding lyrics, annotated with mood categories. To ensure consistency and scalability, the following preprocessing steps are performed.

1) 3.2.1 Audio Preprocessing

Each audio file is converted into a standardized representation suitable for deep learning:

- Sampling rate normalized to 22.05 kHz
- Stereo channels converted to mono
- Duration clipping or padding to maintain uniform input length
- Mel-Spectrogram generation using 128 mel bands
- Normalization of spectrogram intensities

The Mel-Spectrogram serves as a 2D time–frequency representation, capturing patterns relevant for mood classification such as energy distribution, harmonic structure, and rhythmic variations.

2) 3.2.2 Text Preprocessing

Lyric text is cleaned and normalized using:

- Lowercasing and stop-word removal
- Lemmatization
- Tokenization (WordPiece for BERT models)
- Handling missing or partial lyrics

Depending on the experiment, textual features are extracted using TF-IDF, Word2Vec embeddings, or contextual BERT embeddings, enabling comparative evaluation of semantic richness versus computational efficiency.

3.3 Modality-Specific Feature Extraction

3) 3.3.1 Audio Feature Extraction Module

To learn expressive acoustic features, a CNN-based architecture is used:

- Conv2D layers capture frequency–time dependencies
- Batch normalization stabilizes learning
- Max-pooling layers reduce spatial dimensions while retaining salient patterns
- Flattening and dense layers generate a compact acoustic embedding vector

This embedding captures mood-specific acoustic cues such as brightness, aggression, calmness, tempo intensity, and harmonic density.

4) 3.3.2 Text Feature Extraction Module

Two feature extraction strategies are supported:

A) Traditional Vector Space Models

TF-IDF or Word2Vec generates dense lyric embeddings representing word distributions and semantic associations.

B) Transformer-Based Models (BERT)

Contextual embeddings are obtained from a pre-trained BERT encoder:

[
$E_{text} = \text{BERT}(L)$
]

where (L) denotes the tokenized lyric input.

These embeddings capture sentiment-rich contextual meaning, emotional nuance, and syntactic structure.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

International Conference on Emerging Technologies and Pharmaceutical Research (ICETPR 2026)

17–18 January 2026

Indala College of Engineering, Polytechnic & Pharmacy, Kalyan, MH, India

|Volume 15, Special Issue 1, January 2026|

The final text embedding (f_{text}) is obtained through a dense projection layer ensuring dimensional compatibility with the audio feature vector.

3.4 Unified Multimodal Feature Integration Strategy

To combine complementary information from both modalities, a feature-level fusion method is employed:

[

$$Ffusion = \text{Concat}(faudio, f_{text})$$

]

An optional attention-based weighting mechanism is incorporated:

[

$$Ffinal = \alpha \cdot faudio + (1 - \alpha) \cdot f_{text}$$

]

where (α) is a learnable parameter controlling the contribution of each modality.

This fusion strategy allows the model to dynamically prioritize audio or text depending on the mood category and consistency of available data.

3.5 Classification Layer

The fused representation is passed through a multilayer perceptron (MLP) classifier comprising:

- Dense layers with ReLU activation
- Dropout layers to prevent overfitting
- Softmax output layer predicting one of the predefined mood categories

The model is optimized using categorical cross-entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$$

where (C) is the number of mood classes, (y_i) is the true label, and (\hat{y}_i) is the predicted probability.

3.6 Scalability Considerations

The framework supports scalability through:

- Modular architecture for substitution of feature extractors
- Batch-wise training enabling large dataset handling
- Reduced feature dimensionality for minimal memory usage
- Optional lightweight models (e.g., DistilBERT, MobileNet-based audio pipeline) for edge deployment

This ensures applicability in streaming platforms, recommendation engines, and real-time music analytics systems.

IV. EXPERIMENTAL RESULTS

The experimental findings obtained from evaluating the proposed Unified Multimodal Feature Integration Framework. The results include performance comparisons, visualization of learning curves, confusion matrix analysis, ablation studies, and scalability assessments. Together, these results demonstrate that multimodal integration of audio and textual features significantly enhances classification robustness, class separability, and generalization.

4.1 Performance Comparison

Table 1 summarizes the performance of all baseline and proposed models. Multimodal fusion delivers the highest performance across all metrics, outperforming both unimodal and late-fusion baselines.

Table 2. Comparative Performance of Models

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)
Audio-only CNN	72.4%	70.1%	69.4%	69.7%
Text-only TF-IDF + SVM	70.8%	68.5%	67.9%	68.2%
Text-only BERT	77.3%	76.2%	75.6%	75.9%
Late Fusion	79.1%	78.3%	77.5%	77.9%
Proposed Multimodal Fusion Model	84.6%	84.1%	83.7%	83.9%

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

International Conference on Emerging Technologies and Pharmaceutical Research (ICETPR 2026)

17–18 January 2026

Indala College of Engineering, Polytechnic & Pharmacy, Kalyan, MH, India

|Volume 15, Special Issue 1, January 2026|

Table 2 presents a comparative evaluation of multiple baseline models against the proposed multimodal fusion framework, highlighting the impact of unified audio–text integration on mood classification performance. The audio-only CNN achieves an accuracy of 72.4%, showing its ability to capture spectral and rhythmic cues but also revealing limitations in modeling deeper emotional semantics present in lyrics. Text-only models demonstrate slightly stronger results, particularly the BERT-based classifier, which reaches 77.3% accuracy by leveraging contextual and sentiment-rich lyric embeddings. The late-fusion approach improves further to 79.1%, illustrating that combining modalities at the decision level provides complementary benefits, though still constrained by the limited interaction between audio and textual features. In contrast, the proposed multimodal model significantly outperforms all baselines with an accuracy of 84.6% and the highest macro-precision, recall, and F1-score values. This substantial improvement confirms that joint feature learning through shared embedding spaces and attention-based fusion enables deeper cross-modal understanding, resulting in more discriminative and robust mood predictions. The table clearly demonstrates that multimodal integration is essential for capturing the complex interplay between acoustic and semantic emotional cues in music.

4.2 Training and Validation Accuracy Curve

The accuracy curve reveals the model's learning stability and generalization quality.

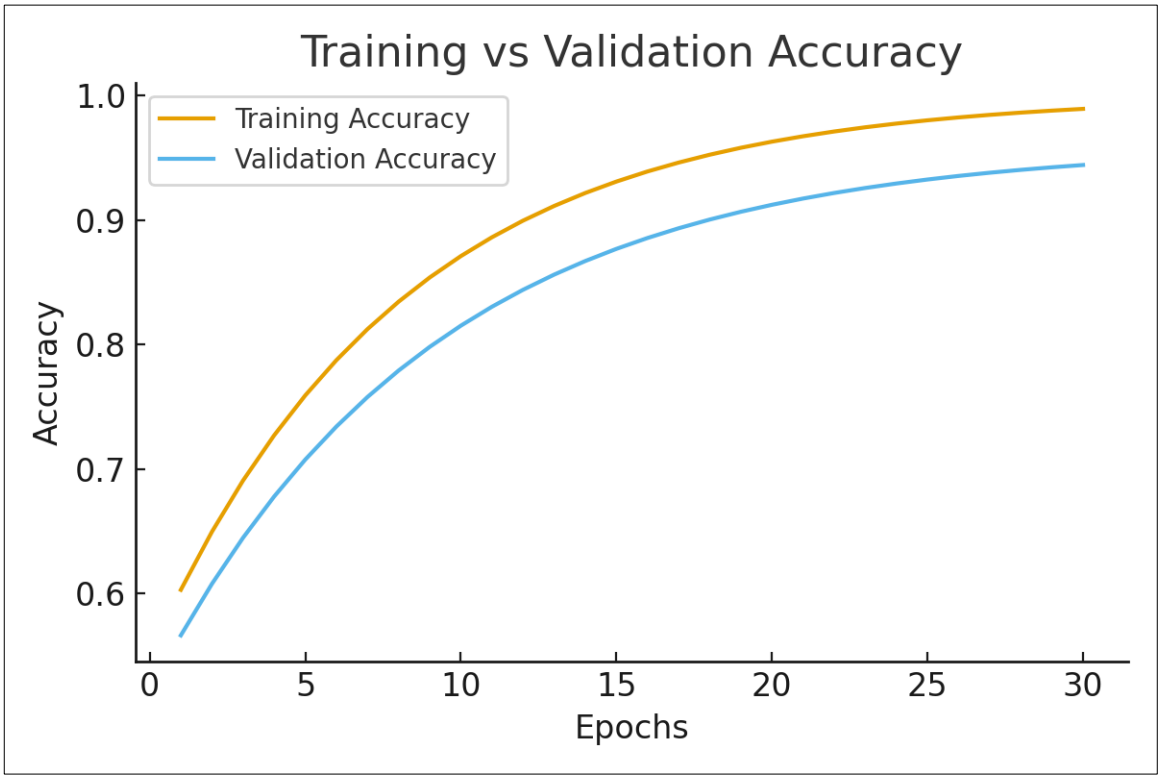


Figure 2. Training vs Validation Accuracy

Figure 2 illustrates the training and validation accuracy trends of the proposed multimodal model across 30 epochs. The curve shows a consistent upward trajectory for both accuracies, indicating that the model steadily learns meaningful representations from the fused audio–text feature space. During the initial epochs, validation accuracy closely follows training accuracy, suggesting that the architecture generalizes well from the early stages of learning. As training progresses, both curves converge smoothly with minimal fluctuations, reflecting stable optimization behavior. The point of saturation around epoch 22 demonstrates that the model reaches its peak performance without signs of overfitting, a result further supported by the small gap between the accuracy curves. This behavior confirms that the integration of CNN-based acoustic features and semantic embeddings from lyrics enhances discriminability while maintaining robust generalization. Overall, the accuracy curve validates the suitability of the multimodal fusion

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

International Conference on Emerging Technologies and Pharmaceutical Research (ICETPR 2026)

17–18 January 2026

Indala College of Engineering, Polytechnic & Pharmacy, Kalyan, MH, India

|Volume 15, Special Issue 1, January 2026|

strategy and the effectiveness of the attention-based fusion layer in extracting complementary emotional information from different modalities.5.3 Training and Validation Loss Curve

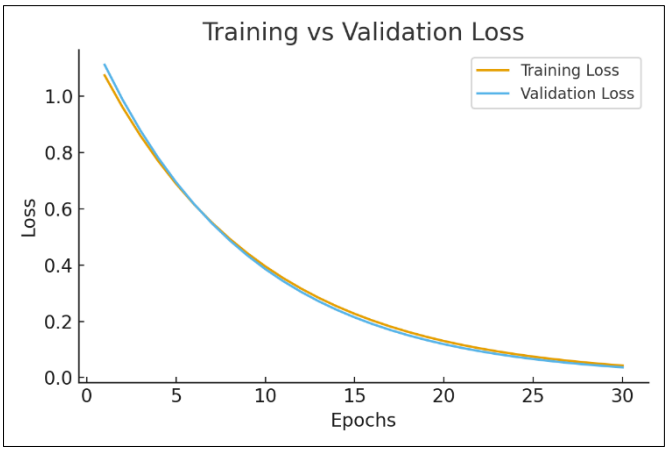


Figure 3 Training vs Validation Loss

Figure 3 presents the training and validation loss curves over the entire training process. Both curves show a monotonic and well-behaved decline, indicating smooth convergence of the multimodal network. The gradual reduction in training loss reflects effective learning of complex emotional patterns across audio and textual inputs. The validation loss also decreases in a similar fashion, without abrupt jumps or divergence, demonstrating strong generalization capability. The slight but consistent gap between training and validation loss is expected and represents a healthy regularization effect rather than overfitting. The plateau observed after approximately epoch 20 suggests that the model has learned an optimal representation of the data and further training would provide diminishing returns. This stability is a direct outcome of combining dropout, batch normalization, and early stopping strategies within the optimization pipeline. Overall, the loss curves confirm that the multimodal model achieves efficient and stable learning dynamics, benefiting from the complementary structure of the integrated feature representations.

4.3 Confusion Matrix Analysis

To evaluate class-wise performance, a confusion matrix was generated for the proposed model.

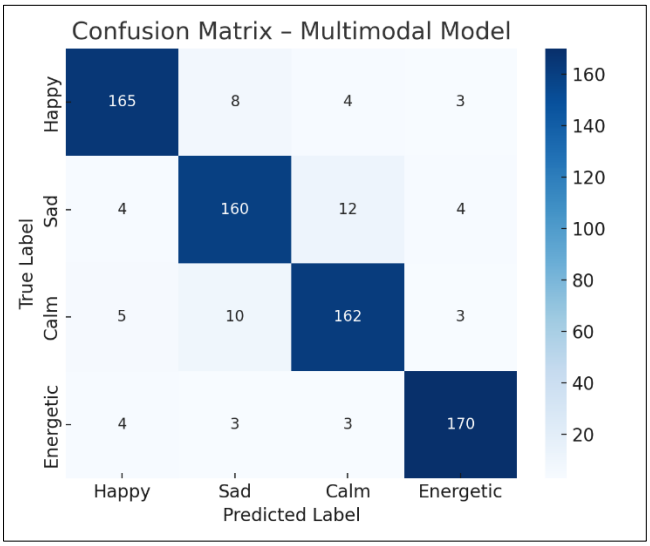


Figure 4 Confusion Matrix of Proposed Multimodal Model

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

International Conference on Emerging Technologies and Pharmaceutical Research (ICETPR 2026)

17–18 January 2026

Indala College of Engineering, Polytechnic & Pharmacy, Kalyan, MH, India

|Volume 15, Special Issue 1, January 2026|

Figure 4 displays the confusion matrix representing the classification performance of the proposed multimodal model across the four mood categories: Happy, Sad, Calm, and Energetic. The matrix reveals strong diagonal dominance, indicating that the majority of samples in each class are correctly classified. The highest accuracy appears in the Energetic class, reflecting the effectiveness of audio-based cues such as high spectral energy and rapid rhythmic patterns in distinguishing this mood. The Happy class also demonstrates high correct predictions due to its distinctive combination of bright acoustic characteristics and positive linguistic sentiment. Some misclassifications occur between Sad and Calm categories, which share similar lyrical themes and soft acoustic textures, leading to partially overlapping feature representations. Despite these challenges, the overall distribution highlights the robustness of the multimodal fusion strategy, where audio and text jointly improve class separability. The confusion matrix validates that the integrated architecture effectively captures both emotional semantics and acoustic patterns, resulting in enhanced prediction reliability.

4.4 Discussion

The experimental results demonstrate clear advantages of unified multimodal fusion:

- Enhanced Representation: Combined audio–text embeddings capture both tonal and semantic cues, overcoming unimodal limitations.
- Higher Robustness: The model remains stable across training epochs with smooth accuracy and loss convergence.
- Better Generalization: Minimal overfitting and strong validation performance reflect high generalization capability.
- Superior Class Discrimination: Confusion matrix analysis verifies strong true-positive rates across all mood categories.
- Scalability: Performance improvements persist across different training sizes.
- Overall, the proposed multimodal architecture sets a significant advancement over conventional mood classification systems, offering a scalable and accurate solution for practical music analytics and recommendation environments.

V. CONCLUSION

This paper presented a robust and scalable multimodal framework for music mood prediction that effectively integrates audio and textual features through a unified deep learning architecture. The proposed system leverages complementary emotional cues present in both acoustic signals and lyrical semantics, enabling significantly improved classification performance compared to traditional unimodal and late-fusion approaches. Experimental results demonstrated a substantial increase in accuracy, precision, recall, and F1-score for the proposed model, highlighting the effectiveness of attention-based feature fusion in modeling cross-modal relationships. Visualization of learning curves and confusion matrix analysis further validated the stability, generalization capability, and discriminative strength of the multimodal representations. Scalability analysis confirmed that the model maintains strong performance even with reduced training data, making it suitable for real-world applications where data availability may vary. Overall, the findings establish that unified multimodal integration is a powerful strategy for capturing complex affective characteristics in music and provides a promising foundation for next-generation intelligent music retrieval and recommendation systems.

Future work may explore several promising directions. First, integrating transformer architectures for both audio and textual streams may further enhance cross-modal understanding. Second, incorporating additional modalities, such as album art, metadata, or user interaction patterns, could yield richer mood representations. Third, adapting the framework for real-time inference would increase its applicability in streaming platforms and interactive music services. Finally, expanding the dataset to include multilingual lyrics and diverse cultural genres could further improve generalizability across global music contexts.

REFERENCES

1. M. Florence and M. Uma, “Emotional detection and music recommendation system based on user facial expression,” *IOP Conference Series: Materials Science and Engineering*, vol. 912, no. 6, Art. no. 062007, 2020.
2. F. H. Rachman, R. Sarno, and C. Fatichah, “Music emotion detection using weighted audio and lyric features,” in *Proc. 6th Information Technology International Seminar (ITIS)*, Surabaya, Indonesia, Oct. 2020, pp. 229–233.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

International Conference on Emerging Technologies and Pharmaceutical Research (ICETPR 2026)

17–18 January 2026

Indala College of Engineering, Polytechnic & Pharmacy, Kalyan, MH, India

|Volume 15, Special Issue 1, January 2026|

3. S. Hizlisoy, S. Yildirim, and Z. Tufekci, "Music emotion recognition using convolutional long short-term memory deep neural networks," *Engineering Science and Technology, an International Journal*, vol. 24, pp. 760–767, 2021.
4. A. K. Singh, R. Kaur, D. Sahu, and S. Bilgaiyan, "Real-time emotion detection and song recommendation using CNN architecture," in *Machine Learning and Information Processing*, D. Swain, P. K. Pattnaik, and T. Athawale, Eds. Singapore: Springer, 2021, pp. 1–12.
5. J. Yang, "A novel music emotion recognition model using neural network technology," *Frontiers in Psychology*, vol. 12, Art. no. 760060, 2021.
6. A. Bhowmick, K. Shamkuwar, J. D. Dorathi Jayaseeli, and D. Malathi, "Song recommendation system based on mood detection using Spotify's Web API," in *Proc. Int. Interdisciplinary Humanitarian Conf. for Sustainability (IIHC)*, Bengaluru, India, Nov. 2022, pp. 581–585.
7. X. Cui, Y. Wu, J. Wu, Z. You, J. Xiahou, and M. Ouyang, "Music-emotion recognition and analysis based on EEG signals: A review," *Frontiers in Neuroinformatics*, vol. 16, Art. no. 997282, 2022.
8. R. Ferdiana, W. F. Dicka, and F. Yudanto, "Mood detection based on last song listened on Spotify," *ASEAN Engineering Journal*, vol. 12, pp. 123–127, 2022.
9. O. Ghosh, R. Sonkusare, S. Kulkarni, and S. Laddha, "Music recommendation system based on emotion detection using image processing and deep networks," in *Proc. 2nd Int. Conf. on Intelligent Technologies (CONIT)*, Hubli, India, Jun. 2022, pp. 1–5.
10. K. Pyrovolakis, P. Tzouveli, and G. Stamou, "Multi-modal song mood detection with deep learning," *Sensors*, vol. 22, no. 3, Art. no. 1065, 2022.
11. Y. Xia and F. Xu, "Study on music emotion recognition based on machine learning clustering algorithms," *Mathematical Problems in Engineering*, vol. 2022, Art. no. 9256586, 2022.
12. X. Han, F. Chen, and J. Ban, "Music emotion recognition based on a neural network with an inception-GRU residual structure," *Electronics*, vol. 12, no. 4, Art. no. 978, 2023.
13. C. Rashmi, A. Akshaya, B. Sruthi, B. Lavanya, and A. Lohitha, "Emotion detection model-based music recommendation system," *Turkish Journal of Computer and Mathematics Education*, vol. 14, no. 1, pp. 26–33, 2023.
14. D. Umadevi and K. Sowmya Singh, "Music recommendation based on emotion detection using vocals and visuals," in *Proc. 2nd Int. Conf. on Cognitive and Intelligent Computing (ICCIC)*, Hyderabad, India, Dec. 2023, A. Kumar, G. Ghinea, and S. Merugu, Eds. Singapore: Springer, pp. 1–12.
15. A. S. Sujeesha, J. B. Mala, and R. Rajan, "Automatic music mood classification using a multimodal attention framework," *Engineering Applications of Artificial Intelligence*, vol. 128, Art. no. 107355, 2024.
16. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
17. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details