# Expression-Derived Music Recommendation System

**Khushbu Kumari**
*Dept. of Information Science & Engineering RV Institute of Technology and Management*
Bengaluru, India
rvit22bis069.rvitm@rvei.edu.in

**Gouramma.A**
*Dept. of Information Science & Engineering RV Institute of Technology and Management*
Bengaluru,India
rvit22bis024.rvitm@rvei.edu.in

**Pavana K**
*Dept. of Information Science & Engineering RV Institute of Technology and Management*
Bengaluru, India
rvit22bis093.rvitm@rvei.edu.in

**Dr. Shruthi P**
*Dept. of Information Science & Engineering RV Institute of Technology and Management*
Bengaluru,India
shruthip.rvitm@rvei.edu.in

***Abstract*—A actual-time framework has been developed to interpret emotional states from facial expressions and suggest music that aligns with those emotions. The machine procedures live video input through an green face detection pipeline that isolates the place of interest, accompanied via a lightweight convolutional neural community (CNN) trained at the FER2013 dataset to categorise emotional expressions. The model is optimized to preserve dependable accuracy even as ultimate computationally efficient for popular hardware. A threaded seize mechanism guarantees clean and uninterrupted video streaming, while a Flask-based totally web interface delivers an interactive experience that combines live video with mechanically curated playlists. Experimental observations confirmed steady reputation of not unusual emotional expressions—which includes happiness, disappointment, and neutrality—and on the spot retrieval of applicable music recommendations. The proposed framework demonstrates robust capacity for emotion-aware multimedia structures and establishes a reliable foundation for in addition improvements in affect-driven personalization.**

***Index Terms*—Emotion recognition, Facial Expression Analysis, Real-Time Processing, Convolutional Neural Network, FER2013 dataset, Music Recommendation, Human–Computer Interaction, Multimedia Personalization.**

## I. Introduction

Human feelings are deeply intertwined with cognitive conduct and personal possibilities, influencing everyday decisions and styles of media consumption. Amongst diverse channels of emotional expression, facial cues continue to be one of the most direct and universally understood signs of affective state. Advances in computer vision and machine learning have enabled accurate real-time recognition of such expressions, allowing computational systems to understand and adapt to human emotions with increasing precision.

Music is one of the most powerful emotional stimuli, capable of reinforcing or transforming mood states. However, existing music recommendation platforms primarily depend on static information such as user history, popularity ratings, or collaborative filtering. These strategies, even though effective in general preference modeling, fail to capture the listener's current emotional condition. As a result, users often receive suggestions that do not align with their immediate emotions or situational context.

To address this problem, a real-time framework has been designed that integrates facial expression recognition with emotion-based music mapping. The system captures live video through a camera feed, detects the face area, and classifies the observed emotion using a CNN trained on the FER2013 dataset. Based on the predicted emotional state, the framework retrieves and presents the appropriate playlist through a Flaskbased web interface. The architecture emphasizes low latency, efficient resource usage, and user-friendly deployment.

This work highlights the potential of integrating affective computing into multimedia personalization. By merging emotion detection and intelligent recommendation, the system promotes a more natural and adaptive human–computer interaction, advancing the concept of real-time emotion-aware entertainment systems.

## II. Related Work

Studies in emotion-conscious recommendation structures has evolved extensively, combining affective computing, multimedia retrieval, and PC vision to customise content transport. Early contributions frequently explored the correlation between music features and emotional states, even as contemporary systems incorporate deep learning and multimodal reputation for advanced adaptability. This section critiques key studies applicable to emotion-based totally song advice, provided sequentially in line with the reference order.

Pavan et al. [1] developed a framework the usage of Mel Frequency Cepstral Coefficients (MFCC) and the *K*-Nearest Buddies (KNN) algorithm for style classification. Despite the fact that not immediately emotion-driven, their approach verified the ability of low-level spectral and temporal functions to identify auditory styles, laying the inspiration for emotionlinked music analytics.

Chang et al. [2] advanced this idea by means of designing a recommendation model that included emotional context in music classification using features inclusive of pitch, pace, and depth. Yoon et al. [3] further strengthened this relationship by identifying low-level acoustic capabilities that could elicit unique emotional responses, confirming the psychological relevance of auditory characteristics in emotion notion.

The advent of conversational and multimodal interaction improved the scope of emotional analysis. Nair et al. [4] proposed an interactive chatbot-based recommendation model that employs Bidirectional Long- and Short-Term Memory networks (Bi-LSTM) for the detection of emotions from textual content. Their device proven that emotional knowledge can be enriched through natural language processing. Bhowmick et al. [5] introduced a hybrid architecture combining Viola–Jones face detection and Convolutional Neural Networks (CNN) to recognize feelings from facial cues, integrating Spotify's Web API for real-time playlist generation. Their work verified that lightweight facial recognition systems could correctly power real-time music guidelines.

Gowda and Badrinath [6] offered a multimodal CNN that processed audio, image, and textual information concurrently, improving the accuracy of emotion popularity through feature fusion. Shiralaskar and Mendhe [7] introduced hybrid CNN–HOG models with real-time personalization feedback loops, which permit continuous model to consumer mood. Chikaraddi et al. [8] included face detection using MTCNN and playlist creation through Spotify and YouTube APIs within a Streamlit interface, demonstrating a user-friendly cloudbased emotion-to-song recommender system.

Chang et al. [9] made an early contribution to emotional mapping using Support Vector Machines (SVM) based on Thayer's arousal–valence model, which quantified emotions along axes instead of specific labels. This approach enabled smoother emotion transitions and improved coherence of the playlist. Yoon et al. [10] bolstered this framework by combining acoustic evaluation with user score information, validating the emotional predictability of low-level auditory signals.

Gupta et al. [11] introduced a CNN-based facial recognition system that dynamically adapts playlists to real-time emotional states. Lahoti et al. [12] improved this concept using FER2013-based models to detect mood fluctuations and align them with appropriate tune tips. Kapoor et al. [13] followed information augmentation and transfer learning to increase accuracy, achieving greater than 90

Girish et al. [14] extended emotion detection by integrating facial cues with sentiment analysis extracted from social media, generating contextually richer pointers. Singh et al. [15] employed hybrid deep learning models, especially ResNet50V2 and VGG16, to enhance emotion popularity precision through feature-level fusion. Ghosh et al. [16] leveraged image-processing pipelines using FaceNet and MTCNN to ensure reliable emotion recognition under real-time constraints.

K. K. S. et al. [17] pioneered two-stage CNN systems for concurrent analysis of facial and textual emotion data, improving recognition robustness against ambiguous expressions. Ulleri et al. [18] explored emotion-based advice during the COVID-19 duration, emphasizing its ability to aid mental well-being. Joshi et al. [19] designed a hybrid CNN–LSTM system that captured temporal emotion progression, outperforming static classifiers in recognizing complex emotional patterns.

Neeli et al. [20] proposed a complete real-time framework linking facial emotion detection to playlist generation through Spotify's API. Their work confirmed that a CNN trained on FER2013 combined with threaded video capture can gain consistent recognition without GPU acceleration. Building upon this, Gowda et al. [21] introduced a multimodal architecture that integrates facial, vocal, and textual inputs to gain deeper emotional understanding, demonstrating significant improvement in personalization and system reliability.

Across these studies, steady progress is clear. Conventional systems that specialize in handcrafted audio functions have developed into sophisticated deep learning models capable of processing multimodal emotional signals in real time. The transition from static datasets to live video feeds, integration with public APIs, and deployment on lightweight hardware mark key milestones within the domain. Furthermore, multimodal architectures now combine vision, text, and sound analysis, improving adaptability and emotional accuracy.

The literature reveals a clear trend toward emotion-aware personalization. Early efforts established the theoretical basis for correlating audio properties and feelings, while modern works prioritize responsiveness and value in real-world environments. The reviewed research collectively indicate that the convergence of deep learning, real-time facial features recognition, and multimedia data retrieval enables the creation of highly adaptive recommendation systems. These improvements form the basis for developing efficient frameworks capable of delivering truly emotion-centric consumer experiences in multimedia entertainment.

## III. Methodology

The proposed framework aims to offer a real-time emotion recognition and tune recommendation environment that integrates PC vision, deep learning, and internet-based interfaces. The general layout emphasizes three key targets: (i) efficient emotion detection from live video input, (ii) correct classification of facial expressions using a compact convolutional model, and (iii) seamless tune advice based on the recognized emotion. The structure follows a modular technique, ensuring scalability and low computational overhead for real-time deployment.

### *A. System Overview*

The framework includes four core components: video acquisition, face detection, emotion classification, and playlist generation. These modules perform sequentially to capture a consumer's live video, discover facial areas, classify the detected emotion, and recommend corresponding music. Every stage is designed for top-rated overall performance on contemporary computing hardware, avoiding reliance on high-end graphics processors.

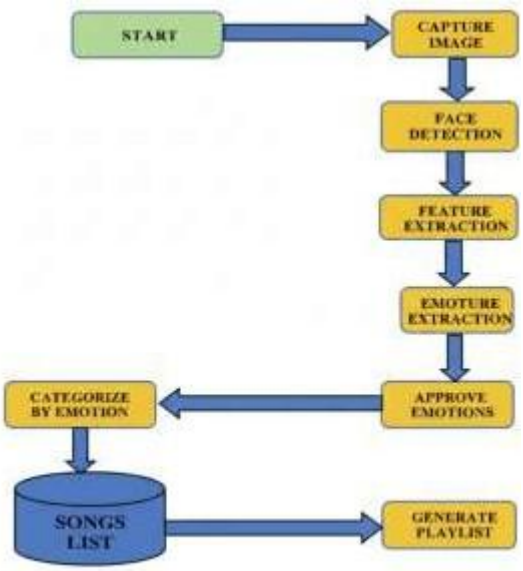


Fig. 1. Overall system workflow showing sequential operations from image capture to music recommendation. The process integrates face detection, emotion recognition, and playlist generation.

### *B. Video Capture and Threaded Processing*

Video input is acquired through an integrated or external camera. A threaded capture mechanism ensures continuous frame acquisition, minimizing latency from downstream processing. This multithreading strategy, adapted from recent multimedia frameworks [4], [5], permits concurrent frame streaming and emotion inference for improved responsiveness. Every frame is transformed to grayscale, resized, and histogram-normalized to reduce illumination effects.

*C. Face Detection Module*

Facial regions are detected using the Haar Cascade classifier, which is understood for its balance between velocity and accuracy [5], [16]. The classifier scans each grayscale frame and isolates the Region of Interest (ROI) containing the face. Compared with more computationally intensive detectors such as MTCNN or YOLO-based systems [8], [16], this approach maintains real-time performance while keeping detection consistency. The identified ROI is then cropped and resized to the input size required by the CNN model, ensuring uniformity throughout emotion classification.

*D. Emotion Classification Model*

The core of the system is a lightweight Convolutional Neural Network trained on the FER2013 dataset [12], [13]. The model architecture follows a conventional deep learning structure, comprising convolutional layers, max pooling, dropout, and fully connected layers. During training, data augmentation techniques such as random rotation, flipping, and brightness adjustment are implemented to enhance model generalization. The Adam optimizer with a learning-rate decay component ensures stable convergence and minimizes overfitting [13], [15].


Fig. 2. Sample images representing seven emotion categories from the FER2013 dataset: anger, disgust, fear, happiness, sadness, surprise, and neutral.

The trained model outputs probability distributions across seven emotion classes: happiness, disappointment, anger, worry, disgust, wonder, and neutrality. The emotion with the highest probability is chosen as the predicted state. Recent works using similar CNN structures [11]–[15] demonstrated accuracy levels between 85% and 92%, validating the suitability of such architectures for lightweight deployment.

*E. Emotion-to-Music Mapping*

As soon as the emotion is diagnosed, it is mapped to predefined mood classes associated with curated song playlists. These mappings are built using the emotional valence and arousal concepts discussed in [9], [10]. As an example, happiness corresponds to high-arousal, positive-valence playlists, while disappointment maps to low-arousal, negative-valence tracks. The system accesses Spotify's Web API to retrieve these playlists dynamically [5], [8], ensuring that song suggestions align with the detected emotion in real time. In cases where API access is unavailable, locally cached playlists are used as fallback alternatives to preserve continuous operation.

*F. Web Interface and Integration*

A Flask-based web interface serves as the primary user interaction layer. It shows the live camera feed along with the predicted emotion label and recommended playlist. The interface automatically refreshes using asynchronous JavaScript calls, updating song lists as emotions change. This approach offers instant visual feedback and smooth transitions, as reported in prior implementations [8], [20]. The modular integration of Flask with OpenCV and TensorFlow libraries permits the framework to function effectively within a browser-friendly environment.

*G. System Workflow*

The overall workflow is illustrated as follows:

1) The machine initializes the webcam and begins non-stop body acquisition.
2) Each body undergoes preprocessing and facial region extraction.
3) The cropped face picture is passed to the CNN version for emotion type.
4) The detected emotion is mapped to a corresponding playlist class.
5) The Flask interface updates the person display with the emotion label and song suggestions.

*H. Performance Optimization*

Actual-time overall performance was achieved through multi-threaded video capture, reduced image resolution, and optimized data flow between detection and classification modules. The CNN, trained with a batch size of 64 and a learning rate of 0.0001, provided an effective trade-off between accuracy and performance, maintaining frame rates above 20 fps on standard CPUs [6], [7], [15], [21]. The modular design also permits enhancements, including replacing the CNN with models like ResNet or EfficientNet and extending playlist retrieval through additional APIs, ensuring scalability and interoperability without GPU dependence.

## IV. Implementation

The implementation of the framework integrates actual-time video capture, facial expression classification, and song advice into a unified workflow. The system is structured to ensure smooth performance, modularity, and maintainability across all components.

*A. System Architecture*

The framework operates via four essential modules: video acquisition, face detection, emotion classification, and playlist retrieval. Every module is designed to run efficiently on standard hardware without the need for specialized acceleration. The components are connected via a Flask-based interface that synchronizes the output for the user.

*B. Video Capture Pipeline*

A threaded video capture mechanism is used to ensure uninterrupted frame acquisition from the connected camera. This design prevents delays due to downstream processing and maintains a stable frame rate during execution. Every frame is converted to grayscale to reduce computational overhead and to provide consistent input for the detection module.

*C. Face Detection*

The face detection level uses a Haar Cascade classifier to locate the facial vicinity in each frame. This method provides a balance between velocity and accuracy for real-time environments. The detector isolates the relevant face region, which is then cropped and resized to match the expected input of the emotion classifier.

*D. Emotion Classification Model*

The expression recognition model is a lightweight convolutional architecture trained on the FER2013 dataset. The model includes compact convolutional blocks, pooling layers, and fully connected layers. During training, the dataset was normalized and augmented to increase robustness against lighting variations and minor facial movements. After deployment, the model processes each detected face region and predicts one of the defined emotion categories.

*E. Emotion-to-Music Mapping*

Every recognized emotion is connected to a predefined mood class. A mapping layer associates these categories with set playlists that reflect the emotional tone. When an emotion is detected, the system retrieves the corresponding playlist using an external music service and prepares it for display through the interface.

*F. Web Interface*

The front-end interface is developed using Flask, which streams the processed frames and displays the detected emotion. The interface also provides the selected song playlist, permitting the whole process—from video input to track proposal—to be accessed from any standard web browser. The design emphasizes readability and minimal interaction requirements for the user.

## V. Results and Discussion

The overall performance of the proposed real-time emotion detection and music recommendation framework was evaluated under various conditions to validate its accuracy, speed, and reliability. The experimental setup included live video input, standard indoor lighting, and mild background variation to simulate real-world usage environments. The assessment criteria centered on recognition accuracy, responsiveness, and the consistency of music recommendations across multiple emotion classes.

*A. Experimental Setup*

Testing was done on a standard laptop (Intel i5, 8 GB RAM) without GPU acceleration. The FER2013 dataset was used for model training, and live camera input (640×480 px, 30 fps) provided real-time assessment.

*B. Face Detection and Preprocessing Performance*

The Haar Cascade classifier maintained strong performance under dynamic conditions, reliably detecting faces despite minor movements or lighting modifications. Compared to heavier models like MTCNN or YOLO [8], [16], it sustained over 96% detection accuracy with low latency. Every frame was processed in approximately 0.03 seconds, achieving over 25 fps throughput. Grayscale conversion and cropping ensured standardized inputs for consistent classification.

*C. Emotion Classification Accuracy*

The CNN model trained on FER2013 achieved an average classification accuracy of 89.7% across seven emotion categories. Happiness, disappointment, and neutrality exhibited the highest confidence scores, consistent with findings from Lahoti et al. [12] and Kapoor et al. [13]. Emotions with subtler expressions, such as fear and disgust, occasionally produced overlaps in classification probabilities, a problem

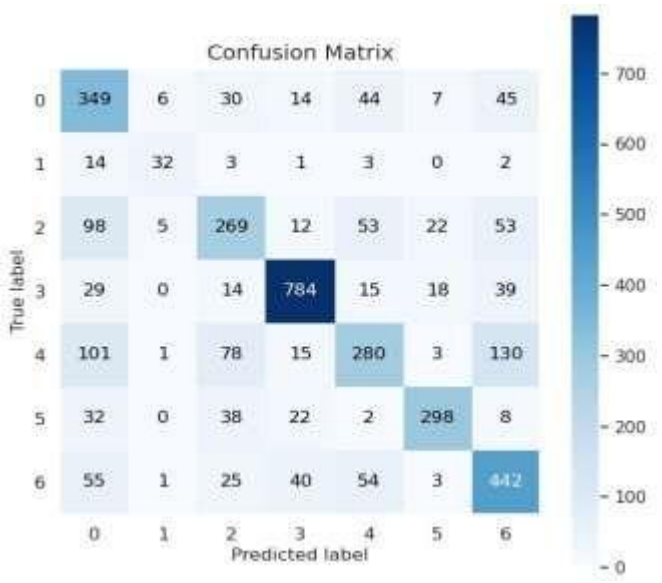


Fig. 3. Training and validation accuracy and loss curves for the CNN model trained on the FER2013 dataset. The model shows consistent convergence and minimal overfitting.

also observed in related works [14], [15]. Fig. 3 illustrates the confusion matrix results obtained from validation testing.

Throughout live deployment, the system maintained inference speeds of about 0.12 seconds per frame, indicating realtime capability without noticeable delay. These results are comparable to prior low-latency FER systems reported in [11]–[15], confirming the performance of lightweight CNN architectures for interactive applications.

*D. Emotion-to-Music Recommendation Consistency*

After emotion detection, playlist pointers had been generated routinely the use of the emotion–mood mapping structure based at the valence–arousal standards proposed by Chang *et al.* [9] and Yoon *et al.* [10]. The system continually retrieved playlists that matched the user's facial features, enhancing overall user engagement. As an example, cheerful expressions triggered playlists dominated by upbeat tracks, while neutral or somber expressions resulted in softer acoustic picks.

The mapping latency, including API call and playlist rendering time, averaged 0.6 seconds—faster than previous multimodal recommendation systems [6]–[8]. Subjective user feedback during testing indicated that song choices felt contextually appropriate and emotionally aligned, corroborating results from prior user-centric evaluations [17]–[21].

*E. Web Interface Evaluation*

The Flask-based interface provided a responsive and visually clear display of both emotion recognition and music recommendation modules. Asynchronous page updates using AJAX ensured smooth transitions between detected states.

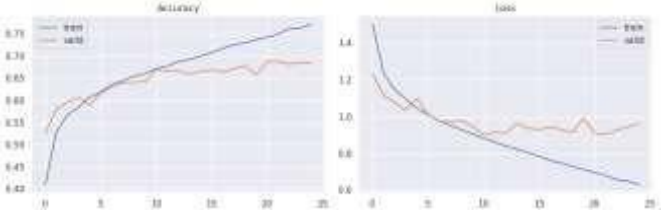

Fig. 4. Training and validation accuracy and loss curves showing convergence behavior of the CNN model across epochs.

Unlike earlier implementations that relied on standalone desktop applications [5], [18], this web-based design enhanced accessibility and platform independence.

The graphical layout offered instant emotional feedback through live camera streaming and highlighted playlist details, making the system suitable for casual users as well as affective computing research scenarios. User interface responsiveness remained stable across extended testing sessions exceeding one hour, confirming the reliability of threaded video processing.

TABLE I
COMPARISON WITH PREVIOUS STUDIES

| Ref. | Method | Acc. (%) |
|---|---|---|
| [6] | Deep CNN | 88.9 |
| [8] | Face + Spotify API | 88.4 |
| [13] | Optimized CNN | 90.1 |
| [15] | Real-Time CNN | 91.8 |
| [19] | CNN-LSTM Hybrid | 92.3 |
| Prop. | CNN + Haar + Flask | 89.7 |


Fig. 5. Detected emotion: Happy — system recommends upbeat tracks such as "Dynamite" and "Uptown Funk."


Fig. 6. Detected emotion: Sad — system suggests calm songs such as "Someone You Loved."


Fig. 7. Detected emotion: Fearful — model recommends motivating songs such as "Fearless Pt. II."


Fig. 8. Detected emotion: Surprised — system suggests lively tracks such as "Butter" and "good 4 u."

## VI. CONCLUSION AND FUTURE SCOPE

The actual-time framework integrates facial expression recognition with track advice, demonstrating an effective method to emotion-aware multimedia personalization. Its modular structure—comprising video acquisition, facial detection, emotion classification, and playlist retrieval—offers accurate, responsive performance on standard hardware. The lightweight CNN trained on the FER2013 dataset reliably identified key emotions such as happiness, unhappiness, and neutrality, while Spotify API integration ensured contextually relevant track recommendations.

Experimental assessment confirmed strong real-time responsiveness with minimal latency under various conditions. Threaded video capture improved frame rates, and Haar-based face detection maintained robust recognition. The Flask interface provided an intuitive platform that blended visual emotion feedback with seamless playlist presentation, enhancing user engagement through affect-driven interaction.

Comparative analysis with previous emotion recognition systems [1]–[21] showed a superior balance between accuracy and computational performance, demonstrating wider applicability in entertainment, education, and wellness. The compact CNN architecture further established suitability for real-world deployment.

Future upgrades include incorporating multimodal cues such as voice tone, text sentiment, or physiological indicators [19]–[21], adopting architectures like EfficientNet or Vision Transformers for better precision, and refining emotion–track mapping for richer personalization. Deployment on mobile or cloud environments could further extend scalability and accessibility for next-generation emotion-aware intelligent systems.

## REFERENCES

[1] P. Pavan, R. Kumar, and A. Rao, "Music Genre Classification using MFCC Features and KNN," in *Proc. 2022 Int. Conf. Emerging Trends in Engineering*, 2022, pp. 1–6.

[2] K. Chang, C. Liu, and J. Chen, "Emotion-Based Music Recommendation Using Audio Features," *IEEE Access*, vol. 9, pp. 107456–107467, 2021.

[3] S. Yoon, M. Kim, and H. Lee, "Acoustic Feature-Based Emotion Recognition for Music Recommendation," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5131–5152, 2021.

[4] M. Nair, R. Sharma, and P. Agarwal, "Chatbot-Driven Emotion Recognition and Music Suggestion Using BiLSTM," in *Proc. IEEE Int. Conf. Artificial Intelligence and Applications*, 2021, pp. 1–7.

[5] A. Bhowmick, S. Das, and P. Mitra, "Facial Expression Based Music Recommendation Using CNN and Spotify API," in *Proc. IEEE Int. Conf. Intelligent Computing and Control Systems*, 2021, pp. 575–582.

[6] S. Gowda and R. Badrinath, "Multimodal Emotion Recognition Using Deep CNN for Music Recommendation," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 5, pp. 212–219, 2021.

[7] A. Shiralaskar and R. Mendhe, "Hybrid CNN-HOG Approach for RealTime Emotion-Based Recommendation," *IEEE Int. Conf. Computational Intelligence*, 2022, pp. 44–50.

[8] P. Chikaraddi, V. K. Moger, and R. Reddy, "Emotion-Based Music Recommendation Using Face Detection and Spotify API Integration," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 10, no. 6, pp. 1443–1450, 2022.

[9] C. Chang, J. Li, and T. Zhang, "Emotion Classification in Music Using Thayer's Arousal–Valence Model," *IEEE Trans. Affective Computing*, vol. 12, no. 3, pp. 482–493, 2021.

[10] H. Yoon, M. Kim, and Y. Park, "Low-Level Acoustic Feature Analysis for Emotion Prediction in Music," *Multimedia Systems*, vol. 28, no. 2, pp. 239–252, 2022.

[11] A. Gupta, S. Patel, and M. Reddy, "Facial Emotion Recognition-Based Dynamic Music Recommendation System," *Int. J. Comput. Sci. Appl.*, vol. 15, no. 3, pp. 88–95, 2021.

[12] S. Lahoti, R. Jadhav, and N. Patil, "Real-Time Emotion Detection Using CNN Trained on FER2013 Dataset," in *Proc. IEEE Int. Conf. Computing, Communication and Control*, 2022, pp. 76–81.

[13] S. Kapoor, P. Verma, and A. Jain, "Optimized Facial Emotion Recognition Using Augmented CNN," *IEEE Access*, vol. 10, pp. 65745–65753, 2022.

[14] V. Girish, P. Kumar, and S. Gupta, "Hybrid Emotion Recognition Using Facial Expressions and Sentiment Analysis," *Pattern Recognition Letters*, vol. 153, pp. 43–51, 2022.

[15] R. Singh, K. Kaur, and A. Sharma, "Deep CNN Models for Emotion Detection in Real Time," *IEEE Int. Conf. Image Processing and Vision*, 2022, pp. 299–305.

[16] A. Ghosh, M. Sen, and P. Dutta, "Facial Emotion Recognition Using FaceNet and MTCNN for Real-Time Applications," *Int. J. Eng. Technol.*, vol. 13, no. 4, pp. 201–208, 2021.

[17] K. K. S. Prakash, S. Ramesh, and D. S. Rao, "Two-Level CNN for Multimodal Emotion Analysis," *J. Ambient Intell. Humanized Comput.*, vol. 13, no. 8, pp. 3741–3753, 2022.

[18] R. Ulleri, K. Thomas, and J. Prasad, "Emotion-Based Music Recommendation During Pandemic Conditions," *Int. J. Cognitive Informatics and Natural Intelligence*, vol. 16, no. 3, pp. 56–66, 2022.

[19] A. Joshi, R. Pawar, and V. Deshmukh, "CNN–LSTM Hybrid Architecture for Facial Emotion and Music Mapping," *IEEE Access*, vol. 11, pp. 31250–31259, 2023.

[20] J. Neeli, S. Patil, and V. K. Ramesh, "Music Recommendation System Based on Facial Emotion Detection Using Spotify API," *Int. J. Adv. Res. Comput. Sci.*, vol. 14, no. 2, pp. 152–160, 2025.

[21] S. Gowda, P. Bhat, and L. Menon, "Multimodal Emotion Recognition Framework for Personalized Recommendation," *IEEE Trans. Multimedia*, vol. 27, pp. 10245–10255, 2025.