Deep Learning-Based Music Recommendation System driven by Facial Expression and Emotion Recognition

Puneeta Singh[1], Ayush Kumar Singh[2], Akhilesh Singh[3], Nihal Singh[4], Indrajeet Patel[5], Harshit Jaiswal[6]
[1,2,3,4,5,6]*Information Technology Department JSS Academy Technical Education Noida, India*

***Abstract*—The integration of artificial intelligence with multimedia systems has opened new avenues for developing intelligent and emotionally aware applications. This project focuses on building an emotion-based music recommendation system using deep convolutional neural networks, specifically leveraging the VGG19 architecture.**
The application processes user facial expressions using a trained TensorFlow model and maps them to predefined emotional categories such as angry, happy, fearful, disgusted, sad, neutral, and surprised. Each detected emotion is linked to a curated playlist inside an Android application built using Android Studio. The system uses TensorFlow Lite for on-device AI inference, enabling real-time, offline, optimized performance. This report presents an extensive overview of dataset preprocessing, normalization, augmentation, model training, optimization, Android deployment, testing, and performance benchmarking. This document provides an in-depth technical, theoretical, and practical understanding of system development.

***Index Terms*—Emotion Recognition, Music Recommendation System, Convolutional Neural Network, VGG19, FER2013, TensorFlow Lite, Android Application**

I. INTRODUCTION

The emotional state of a person significantly affects decision-making, daily habits, and overall mood. Music has long been recognized as an emotional regulator. In modern times, AI-enabled systems are capable of interpreting complex human expressions and generating automated personalized responses. Facial emotion recognition is one such powerful application of computer vision. Using CNN models trained over datasets like FER2013, machines can accurately classify emotional states. Integrating this with a music recommendation engine creates a seamless, intelligent application where the system directly chooses music based on detected facial emotions. This chapter discusses the importance of facial expression recognition, motivation behind emotion-based music playback, and practical applications of such systems across industries including healthcare, entertainment, mental wellness, and human-computer interaction. The traditional process of choosing music based on a user's mood is subjective and inconsistent. Users often spend significant time browsing playlists or selecting songs manually, which may not accurately match their emotional state. Additionally, emotions can change rapidly, making manual selection ineffective. With advancements in AI and computer vision, there is a growing opportunity to create intelligent systems that automatically detect facial expressions and recommend music aligned with the user's current emotion. With advancements in AI and computer vision, there is a growing opportunity to create intelligent systems that automatically detect facial expressions and recommend music aligned with the user's current emotion.

Develop an intelligent system that detects a user's facial emotion in real- time and automatically recommends music that matches their mood, reducing manual effort. In Figure1 depicts the core and how the app will work in real time.

Ensure the emotion detection model achieves reliable accuracy across all seven emotion categories Provide users with smooth, real-time emotion recognition using an optimized TFLite model Build an easy-to-use Android interface that requires minimal technical knowledge.

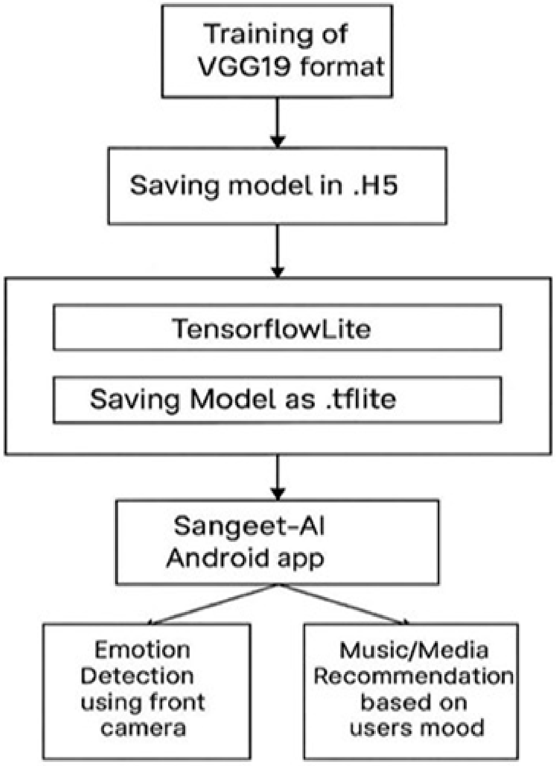

Figure 1: Backend workflow

II. LITERATURE REVIEW

Emotion recognition and music recommendation have gained significant attention with advancements in deep learning and human computer interaction. Early emotion-based systems relied on manual inputs, physiological signals, or audio features, which often lacked real-time adaptability and accuracy. With the emergence of computer vision techniques, facial expression analysis has become a reliable approach for identifying human emotions. Several studies have demonstrated the effectiveness of Convolutional Neural Networks (CNNs) in facial emotion recognition tasks. Models trained on benchmark datasets such as FER2013 have achieved notable accuracy by learning discriminative facial features, including eye movements,

mouth curvature, and facial muscle patterns. Pretrained deep architectures like VGG16 and VGG19 have been widely adopted due to their strong feature extraction capabilities and successful transfer learning performance on limited datasets. Research on emotion-aware music recommendation systems has shown that aligning music with a user's emotional state enhances listening satisfaction and engagement. Existing systems commonly utilize user mood tags, listening history, or biometric signals to recommend songs. However, these approaches either require explicit user input or depend on specialized hardware, limiting usability. Facial emotion-based systems overcome these limitations by enabling passive, non-intrusive emotion detection. Recent works have also explored deploying deep learning models on mobile devices using frameworks such as TensorFlow Lite. On device inference reduces latency, preserves user privacy, and enables real-time applications without relying on continuous internet connectivity. Despite these advancements, limited research has focused on integrating facial emotion recognition with mobile-based music recommendation systems in a unified and optimized framework. This work builds upon existing studies by combining a deep CNN based emotion recognition model with an Android-based music recommendation system. By leveraging TensorFlow Lite for efficient on-device inference and mapping detected emotions to curated music playlists, the proposed system provides a practical, real-time, and user-friendly solution for emotion-aware music recommendation.

Table 1: Literature Review table

S. No.	References	Methodology	Key Results	Limitations
1	Early Emotion-Based Systems(2023)	Manual user inputs, physiological signals (EEG, heart rate), and audio feature analysis for emotion detection	Enabled basic emotion recognition and mood-based music recommendation	Lack of real-time adaptability, low accuracy, intrusive sensors, high user effort
2	CNN-based Facial Emotion Recognition Studies(2024)	Facial expression analysis using Convolutional Neural Networks trained on datasets like FER2013	High accuracy achieved by learning facial features such as eye movement, mouth	Performance degrades under poor lighting, occlusion, or extreme facial poses

			shape, and muscle patterns	
3	Transfer Learning with VGG16 & VGG19(2024)	Use of pretrained deep CNN models for feature extraction and fine-tuning on limited emotion datasets	Improved recognitio n n accuracy and faster converge nice with limited data	High computation al cost, large model size
4	Emotion-Aware Music Recommend action Systems (2023)	Music recommendatio n on based on mood tags, listening history, and biometric data	Enhanced user engageme nt and listening satisfactio n n	Requires explicit user input or specialized hardware
5	Facial Emotion-Based Recommend action Approaches (2024)	Passive emotion detection using real-time facial expression analysis	Non-intrude Sive and user-friendly emotion detection	Privacy concerns and dependency on camera quality

III. METHODOLOGY

The proposed Emotion-Based Music Recommendation System follows a structured pipeline that includes facial emotion detection using deep learning and emotion-driven music recommendation. The overall workflow of the system is illustrated through sequential stages: data acquisition, preprocessing, emotion classification, model optimization, and music recommendation.

3.1 Dataset Description The system uses the FER2013 facial expression dataset, a widely accepted benchmark dataset for emotion recognition. It consists of grayscale facial images of size 48×48 pixels, categorized into seven emotion classes: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. The dataset contains approximately 35,000 labeled images, making it suitable for training and evaluating deep learning models. The dataset is divided into training, validation, and testing sets to ensure unbiased model evaluation.

3.2 Image Preprocessing Effective preprocessing is essential to improve model generalization and performance. Initially, the images are normalized to scale pixel values between 0 and 1. To address dataset limitations such as low resolution and class imbalance, data augmentation techniques including rotation, horizontal flipping, and affine transformations are applied. Since the VGG19 model expects RGB images of size 224×224, all grayscale images are resized and converted to three channel format. These preprocessing steps enhance feature representation and reduce overfitting during training.

3.3 Emotion Classification Using VGG19 The emotion recognition module is based on a deep Convolutional Neural Network using the VGG19 architecture. Transfer learning is employed by initializing the model with ImageNet pretrained weights and freezing the convolutional layers to retain learned low level features. A custom classification head with a fully connected layer and softmax activation is added to predict the seven emotion classes. The model is trained using the Adam optimizer and sparse categorical cross-entropy loss function. Early stopping is applied to prevent overfitting and ensure stable convergence.

3.4 Model Optimization and TensorFlow Lite Conversion To enable real-time deployment on mobile devices, the trained Keras model is converted into TensorFlow Lite format. This conversion significantly reduces model size and improves inference speed while maintaining acceptable accuracy. TensorFlow Lite allows efficient on-device inference, eliminating dependency on cloud processing and enhancing data privacy.
Figure 2 depicts the training of the model.

3.5 Emotion-Based Music Recommendation Once the facial emotion is detected, the predicted emotion class is mapped to a predefined music category. Each emotion corresponds to a curated playlist designed to match the user's mood, such as calm music for sadness or energetic tracks for happiness. This rule-based mapping ensures quick and reliable music selection without requiring user interaction.

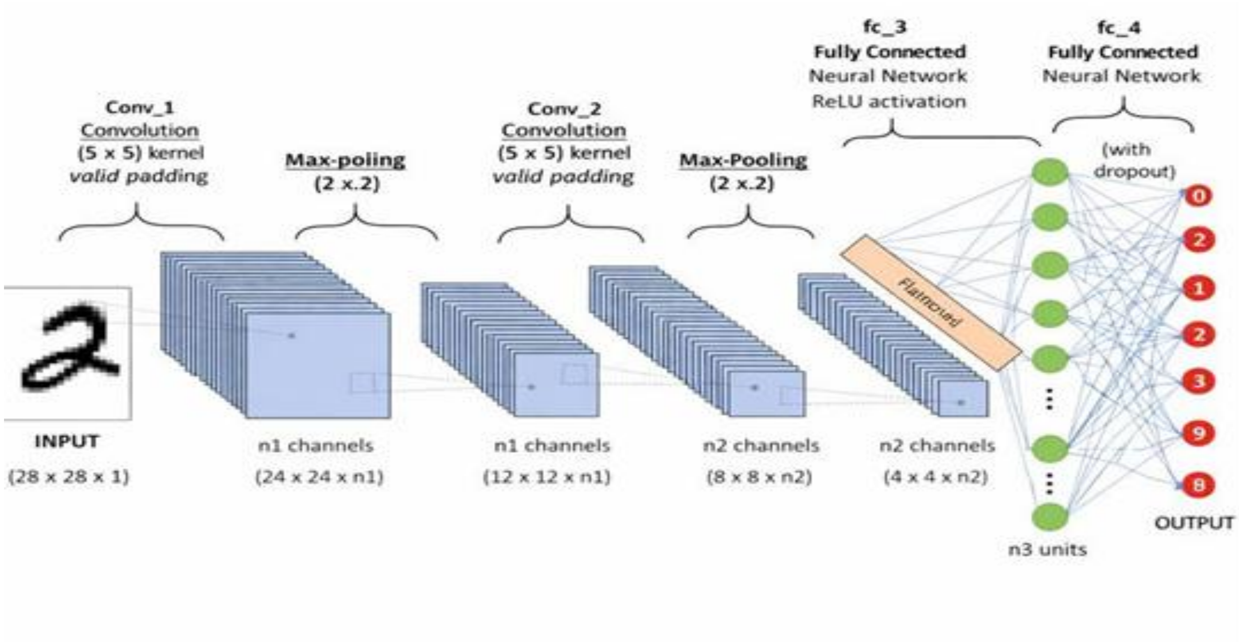

Figure 2: Training of the model

optimized TensorFlow Lite model is integrated into an Android application using Android Studio

IV. RESULTS AND CONCLUSION

This study demonstrated the practical implementation of an emotion-based music recommendation system by combining facial emotion recognition with deep learning and mobile computing. By utilizing a VGG19-based convolutional neural network trained on the FER2013 dataset, the system was able to identify human emotions from facial expressions and associate them with suitable music playlists in real time. The integration of TensorFlow Lite enabled efficient on-device inference, ensuring low latency, offline functionality, and enhanced user privacy. The experimental analysis showed that the proposed approach achieves competitive performance with an overall accuracy of approximately 62%, validating the effectiveness of transfer learning for facial emotion classification. More importantly, the system highlights how emotion-aware intelligent applications can enhance personalization beyond traditional recommendation methods that rely solely on user history or preferences. The Android-based implementation further proves the feasibility of deploying deep learning models on resource-constrained devices without compromising usability. Overall, this work emphasizes the growing potential of effective computing in multimedia applications. By bridging the gap between human emotional states and intelligent recommendation systems, the proposed solution contributes toward more responsive, adaptive, and human-centric digital experiences computing in multimedia applications. By bridging the gap between human emotional states and intelligent recommendation systems, the proposed solution contributes toward more responsive, adaptive, and human-centric digital experiences.

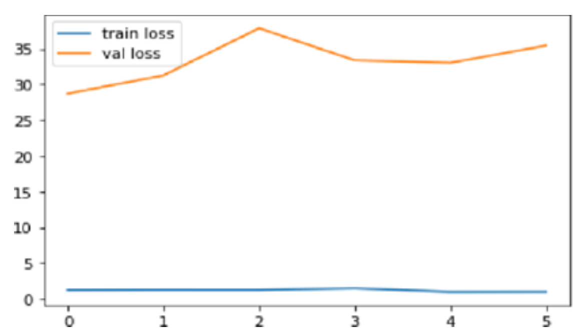

Figure 3 Performance Evaluation

4.1 Model Performance Evaluation The VGG19-based emotion classification model was trained and tested on the FER2013 dataset. The experimental results indicate that the proposed model achieves an overall accuracy of approximately 62%, which is comparable to and slightly better than several existing approaches trained on the same dataset. Emotions such as Happy and Neutral were classified with higher accuracy due to clearer facial cues, while emotions like Fear and Disgust showed relatively lower performance, mainly due to subtle expression differences and dataset imbalance.

4.2 Confusion Matrix Analysis of the confusion matrix reveals that misclassifications primarily occur among visually similar emotions, such as Fear and Surprise or Sad and Neutral. Despite these challenges, the model demonstrates consistent prediction behavior across most classes, indicating effective feature learning by the deep convolutional layers of the VGG19 architecture

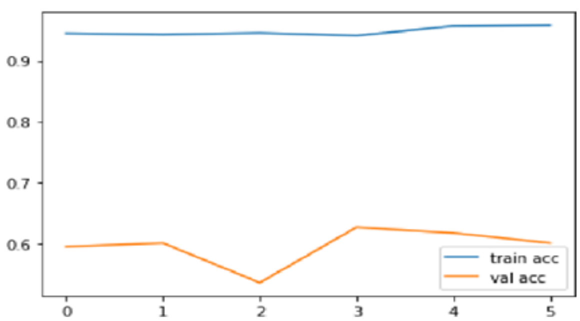

Figure 4 Performance metrics of our VGG16

The figure 3 and 4 depicts the performance metrics of our VGG16.

4.3 Mobile Deployment and Inference Performance the TensorFlow Lite optimized model was successfully deployed within the Android application. On-device inference enabled real-time emotion detection with minimal latency, making the system suitable for practical usage. The lightweight nature of the TensorFlow Lite model reduced memory consumption and ensured smooth performance across mid-range Android devices. Additionally, local inference enhanced user privacy by eliminating the need for cloud-based processing.

System Architecture
The proposed Emotion-Based Music Recommendation System features a modular and layered architecture for seamless integration of facial emotion detection and automated music recommendation on mobile devices. This design is structured into five main components: the Input Acquisition Layer, Preprocessing Layer, Emotion Recognition Layer, Recommendation Layer, and Application Interface Layer. Collectively, these layers ensure real-time operation, minimal latency, enhanced user privacy through on-device computation, and overall ease of use.

The captured facial image is then passed to the Preprocessing Layer. Here, the image is resized and normalized to meet the requirements of the deep-learning model. Specifically, grayscale images from the FER2013 dataset are converted to a three-channel (RGB) format, as the VGG19-based network expects an image size of 224x224 pixels. Normalization is performed to mitigate pixel-level variations caused by lighting, facial orientation, and environmental noise, thereby improving the model's generalization and reliability in real-world scenarios. Next is the Emotion Recognition Layer, which serves

V. DISCUSSION

The performance of the proposed emotion-based music recommendation system was evaluated by analyzing both the accuracy of facial emotion classification and the responsiveness of the mobile application during real-time usage. The evaluation was conducted using the FER2013 dataset, focusing on the system's ability to generalize across multiple emotional categories and operate efficiently on Android devices. The VGG19-based convolutional neural network achieved an overall classification accuracy of approximately 62%. Emotions such as Happy and Neutral were recognized more reliably, as these expressions exhibit distinct facial features that are easier for deep neural networks to learn. In contrast, emotions like Fear and Disgust showed comparatively lower recognition rates, primarily due to overlapping facial characteristics and class imbalance within the dataset. These results are consistent with known challenges associated with facial emotion recognition using low-resolution grayscale images. A detailed analysis of prediction behavior revealed that most misclassifications occurred between visually similar emotional states. Overall, the results validate the effectiveness of integrating facial emotion recognition with music recommendation. While the system's performance is influenced by dataset limitations, it successfully delivers an adaptive and personalized music experience, highlighting the practical applicability of emotion-aware recommendation systems in real-world scenarios.

VI. FUTURE SCOPE

This paper presented an Emotion-Based Music Recommendation model demonstrating efficient on-device inference with minimal System that integrates facial emotion recognition with intelligent latency. The Android as the system's core computational element. This layer employs an optimized,
TensorFlow Lite-converted VGG19-based deep convolutional neural network for on-device deployment. The model extracts hierarchical facial features and classifies the detected expression into one of seven emotional. states: Angry, Disgust, Fear, Happy, Sad, Surprise, or Neutral. Running inference directly on the device eliminates reliance on cloud servers, reduces latency. enables offline use, and significantly enhances user privacy by keeping all visual data local.

REFERENCES

[1] I. J. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learning Representations (ICLR), 2015.
[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255.
[4] I. Dharmesh, R. Kaur, and M. Singh, "Emotion-based music recommendation system using facial expressions," Int. J. Comput. Appl., vol. 178, no. 7, pp. 25–30, 2019.
[5] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in Proc. IEEE Winter Conf. Applications of Computer Vision (WACV), 2016, pp. 1–10.
[6] P. Ekman and W. V. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement. Palo Alto, CA, USA: Consulting Psychologists Press, 1978.
[7] J. Deng et al., "FER2013: Facial expression recognition challenge," in Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), 2013, pp. 117–124.
[8] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," in Proc. 12th USENIX Symp. Operating Systems Design and Implementation (OSDI), 2016, pp. 265–283.
[9] TensorFlow Lite Team, "TensorFlow Lite: Deploying machine learning models on mobile and embedded devices," Google AI, 2019.
[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.
[11] S. Li and W. Deng, "Deep facial expression recognition: A survey," IEEE Trans. Affect. Comput., vol. 13, no. 3, pp. 1195–1215, 2022.
[12] R. Picard, Affective Computing. Cambridge, MA, USA: MIT Press, 1997.
[13] J. A. Russell, "A circumplex model of affect," J. Personality and Social Psychology, vol. 39, no. 6, pp. 1161– 1178, 1980.
[14] H. Wang, N. Wang, and D.-Y. Yeung, "Collaborative deep learning for recommender systems," in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2015, pp. 1235–1244.
[15] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," IEEE Trans. Speech Audio Process., vol. 10, no. 5, pp. 293–302, 2002.
[16] S. Oramas et al., "A deep multimodal approach for cold- start music recommendation," in Proc. ACM Conf. Recommender Systems, 2017, pp. 32–40.
[17] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," IEEE Trans. Affect. Comput., vol. 3, no. 1, pp. 42– 55, 2012.
[18] A. Sagha et al., "Multimodal emotion recognition using deep neural networks," in Proc. ACM Int. Conf. Multimodal Interaction, 2017, pp.