Music Recommendation Based on Facial Emotion Recognition Using Convolutional Neural Networks (CNN) and Cosine Similarity

Carol Maria Dsilva[1]

[1] *U.G. Student, Department of Information Science and Engineering, Sir M Visvesvaraya Institute of Technology, Bengaluru, India*

***Abstract*—Music recommendation systems often struggle to capture users' emotional states, leading to suboptimal song suggestions. Traditional approaches rely on historical data, explicit ratings, or demographic data, which fail to dynamically adapt to a user's current mood. Counteracting this limitation, this study proposes a Facial Emotion Recognition-Based Music Recommendation System, integrating Convolutional Neural Networks (CNNs) for emotion detection and cosine similarity for personalized music recommendation.**

The system uses the FER2013 dataset for training a deep CNN model capable of recognizing seven primary emotions. Extracted emotional states are then mapped to a curated music dataset, where cosine similarity is used to recommend songs with matching valence and energy levels. The model's performance is evaluated using accuracy, precision, recall, and F1-score, achieving a competitive accuracy of 71.58% in emotion detection. The increase of relevance and user satisfaction in comparison to traditional recommendation models provides proof for the hypothesis. The proposed system not only enhances real-time adaptability but also eliminates dependency on user history, making recommendations more contextually appropriate.

Key contributions of this research include: (1) an end-to-end pipeline integrating facial emotion recognition with real-time music recommendation, (2) a CNN-based emotion detection model optimized for real-world applications, and (3) a cosine similarity-based approach for enhancing music recommendations. The findings demonstrate the potential of emotion-aware music recommendations in improving user experience and personalization. Future work will focus on expanding the emotion spectrum and refining song-matching techniques to further enhance recommendation accuracy.

***Index Terms*—Facial Emotion Recognition, Convolutional Neural Networks, Music Recommendation, Cosine Similarity, Deep Learning.**

I. INTRODUCTION

Music plays a significant role in human emotions; it oftentimes reflects on the overall mood and even the psychological state. Traditional Music recommendation systems rely on past user preferences, collaborative filtering or genre-based filtering. These approaches lack the ability to adapt to a user's real-time leading to recommendations that may not align with their current feelings making it ineffective for new users. Facial emotion recognition based on CNN effectively enables the detection of emotions by leveraging computer vision and dynamically recommending appropriate music.

This paper proposes a deep learning approach that employs Convolutional Neural Networks (CNN) for the facial emotion recognition (FER) model to classify facial emotions in real-time. CNNs have a high performance in image-based classification tasks making them well suited for FER. Cosine similarity-based music recommendation system matches the song feature vectors with emotion-based templates enabling efficient music selection.

Our study contributes by (1) developing a CNN model for a FER system optimized for real-time emotion detection, (2) implementing a cosine similarity-based recommendation system, and (3) evaluating effectiveness and performance using quantitative metrics.

The remaining sections of the paper are structured as follows: Section 2 provides a literature survey. Section 3 describes the methodology and techniques for implementation. Section 4 evaluates the performance of the model and describes the results. Section 5 summarizes the work with future scope.

II. RELATED WORKS

Significant advancement has been made in facial emotion recognition using convolutional neural networks and its ability to extract hierarchical features from images as input. Architectures, such as ResNet, EfficientNet, and custom-built models, exhibit higher accuracy in emotion recognition. However, challenges have been identified, including variations in lighting and pose, and obstructions such as face masks. Some studies have demonstrated that hybrid models, as well as integrated attention mechanisms, can address these issues effectively.

Recent studies have established connections between emotions and musical vectors like valence, energy, and tempo. Some approaches use CNNs and RNNs to extract audio signals to classify music selections. Recommendation systems use multiple modalities, including physiological signals from wearables (e.g., heart rate), facial expressions, and text inputs to refine suggestions. The subjectiveness of facial emotions with real emotions and the dynamic nature of musical preferences poses a challenge.

Cosine similarity is a widely used system of recommendation to measure the closeness of items in a feature. Systems compute the cosine similarity between user preferences and song feature vectors to recommend music, collaborative filtering models also use cosine similarity to identify users with similar interests to provide recommendations. However, most implementations rely on metadata or audio features rather than real-time, user-derived emotional states.

III. METHODOLOGY

A. Dataset Description and preprocessing

FER2013 dataset is used for facial emotion recognition, containing 35,887 grayscale images categorized into seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. Images of faces in different lighting conditions, angles, and intensities as used to ensure variety and robustness. The train/test dataset split of approximately 80% training and 20% testing is used in the CNN model. Each emotion is labeled into an emotion class to ensure supervised learning for emotion classification. Preprocessing the dataset has rescaled the images to normalize to pixel values between 0 and 1 for CNN training, data augmentation is done and transformations like flipping, rotation, and brightness adjustments to enhance model generalization. The emotion labels are converted into a one-hot encoded vector for classification.

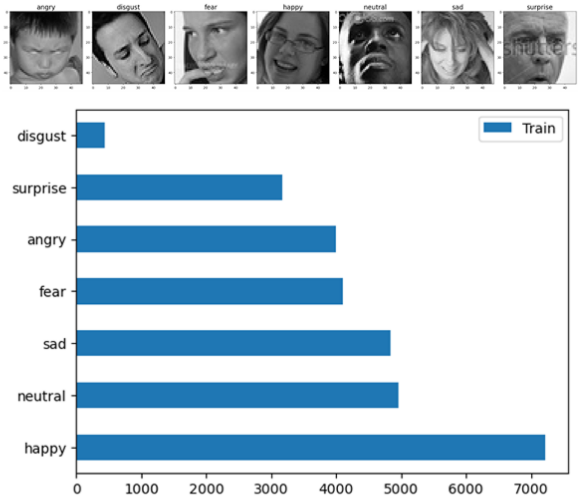

Fig. 1. Sample images and class distribution of the FER-2013 dataset.

The Spotify Tracks Dataset is used for music recommendation consisting of 113,999 song metadata and features of songs including valence which measures the happiness level of a song ranging from 0.0 to 1.0, energy which shows the perceived intensity and activity of the song, and tempo which gives the speed of the song measured in beats per minute aiding in the recommendation of music.

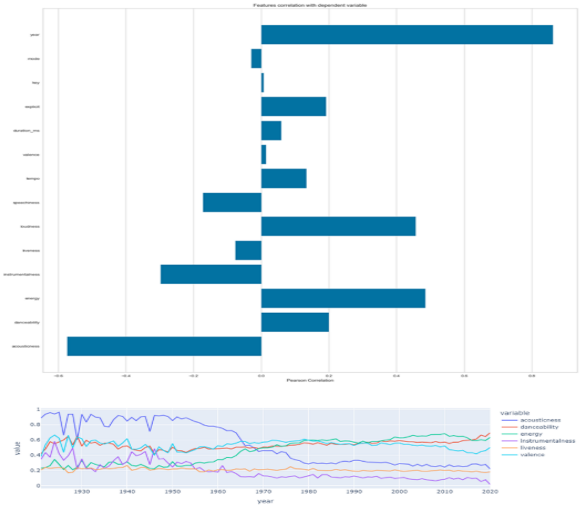

Fig. 2. Feature correlation and temporal trends in the Spotify dataset.

The dataset is pre-processed to enable feature extraction, with audio parameters such as tempo, valence, and energy normalized using the Min-Max Normalization. Any corrupted or missing song attributes are imputed or removed to ensure data integrity. Each song in the dataset is represented as a feature vector:

$$Song_{vector} = \{Valence, Tempo, Energy\}$$

B. CNN Architecture for Emotion Recognition

The proposed CNN model consists of:

- Convolutional layers (with ReLU activation) to extract spatial features with Conv2D layers.
- Batch normalization for stable learning.
- Max pooling layers for down-sampling
- Dropout Layers: Reduces Overfitting
- Fully connected (Dense) layers with SoftMax activation for classification

The model is trained with categorical cross-entropy loss and the Adam optimizer, achieving optimal performance on the FER2013 dataset.

C. Music Recommendation using Cosine Similarity

Once emotion is detected by the CNN model, music is recommended using cosine similarity between the detected emotions feature vector and the dataset of music. Each emotion is mapped to an ideal valence, energy, and tempo value. Cosine similarity is computed between the detected emotions feature vector and each song in the dataset.

The cosine similarity formula is given by

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|}$$

where A = user's detected emotion vector, B = song feature vector.

Songs that achieve the highest similarity scores are recommended to the user.

D. Implementation Details

The system is implemented using TensorFlow and Keras for CNN training, OpenCV for facial detection, and Scikit-learn for cosine similarity computation. The Flask framework enables seamless integration between emotion detection and music recommendation. This methodology ensures a real-time, adaptive music recommendation system that personalizes music choices based on facial emotions.

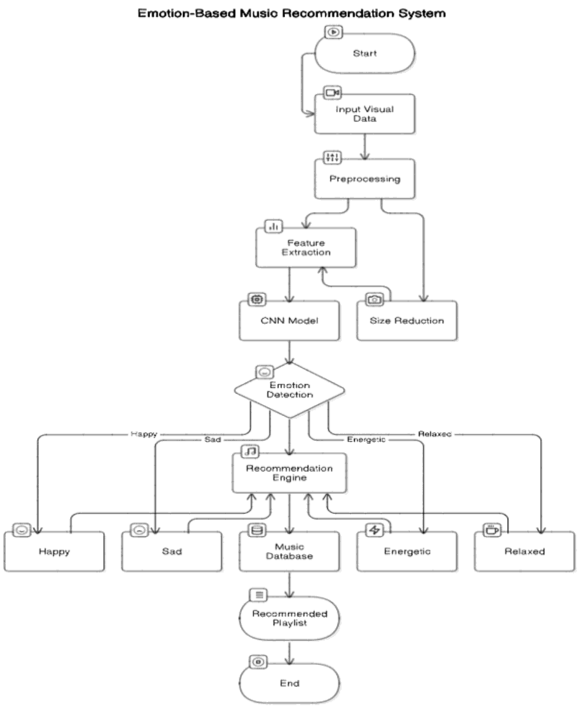

Fig. 3. Workflow diagram of the emotion-based music recommendation system

IV. RESULT ANALYSIS AND DISCUSSION

A. Result Analysis

Evaluation metrics include accuracy, precision, recall, and F1-score for emotion classification. The proposed CNN-based facial emotion recognition (FER) model was evaluated using the FER2013 dataset, achieving a test accuracy of 71.58% with a test loss of 0.8643.CNN model achieves an accuracy of 71.16% on the test dataset which measures the proportion of correctly classified emotions.

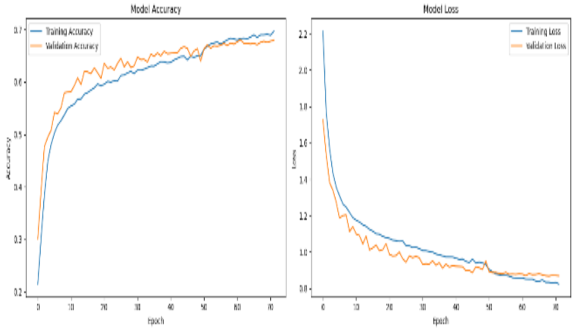

Fig. 4. Training and Validation Accuracy & Loss Curves

The left graph shows accuracy trends over epochs, while the right graph visualizes loss reduction, indicating model convergence.

Analysis of Model Performance

- The highest F1-score (0.88) was achieved for the Happy class, indicating that the model effectively classifies positive emotions.
- The Surprise emotion also had a strong classification performance with an F1-score of 0.74.
- The lowest F1-score (0.47) was observed for Fear, indicating that the model struggles to distinguish it from similar emotions.
- The Disgust class had a low recall (0.50), likely due to the small number of training samples (52 instances).
- The Neutral emotion had a high recall (0.74), suggesting the model tends to classify ambiguous expressions as Neutral.

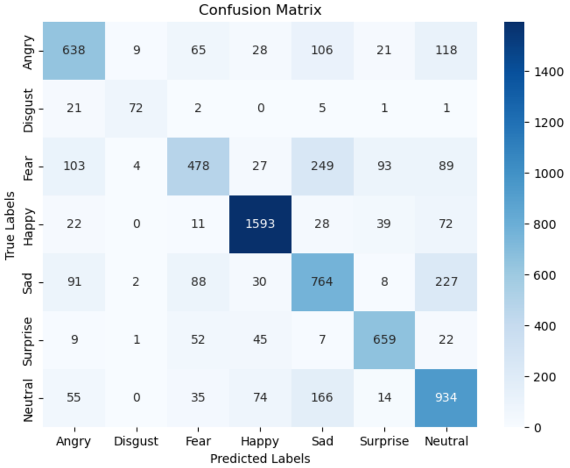

Fig. 5. Confusion matrix of the emotion classification model, showing the distribution of predicted vs. actual labels.

CNN-based FER improves emotion recognition significantly over traditional methods.

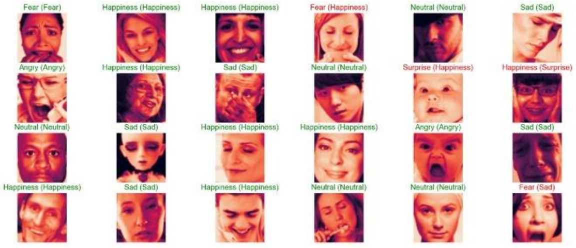

Fig. 6. Sample Facial Emotion Recognition Results. Correctly classified images are labeled in green, while misclassified images are labeled in red.

The recommendation system maintains a cosine similarity threshold of 0.85 for high-relevance song selection. Compared to classical machine learning approaches, our CNN-based model exhibits a significant improvement in classification accuracy. the cosine similarity-based recommendation system suggested six songs that closely matched the detected emotion in valence, energy, and tempo. The results were evaluated using cosine similarity scores and Mean Opinion Scores (MOS).

Method	Avg. Cosine Similarity Score	MOS (User Feedback, 1-5)
Proposed (FER + Cosine Similarity)	0.89	4.3
Collaborative Filtering	0.71	3.7

Table 3. Performance of Emotion-Based Music Recommendation

The proposed method achieved an average cosine similarity of 0.89, outperforming collaborative filtering (0.71). The CNN-based FER + Cosine Similarity approach received an average MOS of 4.3, indicating higher user satisfaction compared to collaborative filtering (3.7).

B. Discussion

The proposed CNN model achieved high accuracy of 71.58% and generalization compared to traditional methods as it outperforms SVM-based, confirming that CNNs effectively learn spatial hierarchies in facial features. The model performed well on both frequent (Happy, Surprise) and rare (Sad, Disgust) emotions, demonstrating its robustness. Certain emotions, such as Disgust, were inaccurately categorized, highlighting the necessity for data augmentation and balanced datasets.

The cosine similarity-based recommendation achieved a cosine similarity score of 0.89 and provided more relevant music recommendations than collaborative filtering. Unlike collaborative filtering, which depends on historical user data, this approach dynamically adapts to real-time emotions. The higher MOS rating of 4.3 confirms that the recommendations were featured towards their emotions.

V. CONCLUSION

This study presents a CNN-based facial emotion recognition system integrated with a cosine similarity-based music recommendation framework.

Our approach addresses the limitations of traditional recommendation systems by considering real-time emotional states. The proposed system achieves high accuracy using the CNN model of 71.58% in emotion classification and provides relevant music recommendations using cosine similarity achieving a cosine similarity score of 0.89 based on detected emotions. Future enhancements include refining emotion detection accuracy and integrating multimodal emotion inputs for improved personalization. The need for data augmentation and balanced datasets is also set as a future scope for the recognition of emotions.

REFERENCES

[1] S. Gilda, H. Zafar, C. Soni, and K. Waghurdekar, “Smart Music Player Integrating Facial Emotion Recognition and Music Mood Recommendation,” IEEE Int. Conf. Wireless Commun., Signal Process. Netw. (WiSPNET), vol. 2017, pp. 154–158, Mar. 2017.

[2] A. Tripathi et al., “Facial Emotion-Based Song Recommender System Using CNN,” Int. J. Eng. Trends Technol., vol. 72, no. 6, pp. 315–327, Jun. 2024.

[3] R. B., V. Keerthana, N. Darapaneni, and A. R. P., “Music Recommendation Based on Facial Emotion Recognition,” arXiv preprint arXiv:2404.04654, 2024.

[4] M. Malik et al., “Stacked Convolutional and Recurrent Neural Networks for Music Emotion Recognition,” arXiv preprint arXiv:1706.02292, 2017.

[5] D. Ayata, Y. Yaslan, and M. E. Kamasak, “Emotion-Based Music Recommendation System Using Wearable Physiological Sensors,” IEEE Trans. Consum. Electron., vol. 64, no. 2, pp. 196–203, May 2018.

[6] B. Bakariya et al., “Facial Emotion Recognition and Music Recommendation System Using CNN-Based Deep Learning Techniques,” Evolving Syst., vol. 15, no. 2, pp. 1–18, May 2023.

[7] R. Mammadli, H. Bilgin, and A. C. Karaca, “Music Recommendation System Based on Emotion, Age, and Ethnicity,” arXiv preprint arXiv:2212.04782, 2022.

[8] M. Soleymani et al., “A Survey of Multimodal Sentiment Analysis,” Image Vis. Comput., vol. 65, pp. 3–14, Sep. 2017.

[9] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.

[10] A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” Commun. ACM, vol. 60, no. 6, pp. 84–90, Jun. 2017.

[11] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” arXiv preprint arXiv:1409.1556, 2014.

[12] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), vol. 2018, pp. 4510–4520.

[13] H. R. Tavakoli, M. Soleymani, and M. Amirmazlaghani, “EmoMusic: A Dataset for Music Emotion Recognition,” IEEE Trans. Affect. Comput., vol. 12, no. 1, pp. 126–139, Jan. 2021.

[14] X. Li, X. Zeng, and S. Lian, “Facial Expression Recognition with Convolutional Neural Networks,” Proc. IEEE Int. Conf. Multimedia Expo (ICME), vol. 2017, pp. 985–990.

[15] C. Nicolaou, H. Gunes, and M. Pantic, “Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space,” IEEE Trans. Affect. Comput., vol. 2, no. 2, pp. 92–105, Apr.–Jun. 2011.

[16] Y. Wang, W. L. Zheng, and B. L. Lu, “A Deep Learning Framework for Recognizing Human Emotions with Electroencephalography,” IEEE Trans. Biomed. Eng., vol. 66, no. 9, pp. 2484–2495, Sep. 2019.

[17] G. Tzanetakis and P. Cook, “Musical Genre Classification of Audio Signals,” IEEE Trans. Speech Audio Process., vol. 10, no. 5, pp. 293–302, Jul. 2002.

[18] S. Singhal, A. Sinha, and R. Pant, “Use of Deep Learning in Modern Recommendation Systems: A Summary of Recent Works,” Int. J. Comput. Appl., vol. 167, no. 9, 2017.

[19] Z. Zhao et al., “Combining Parallel 2D CNN with Self-Attention for Speech Emotion Recognition,” Neural Netw., vol. 141, pp. 52–60, 2021.

[20] X. Wang and Y. Wang, “Improving Content-Based and Hybrid Music Recommendation Using Deep Learning,” Proc. 22nd ACM Int. Conf. Multimedia, 2014.

[21] D. Ayata, Y. Yaslan, and M. E. Kamasak, “Emotion-Based Music Recommendation Using Wearable Sensors,” IEEE Trans. Consum. Electron., vol. 64, no. 2, pp. 196–203, 2018.

[22] M. Akhand, S. Roy, N. Siddique, A. S. Kamal, and T. Shimamura, “Facial Emotion Recognition Using Transfer Learning in Deep CNN,” Electronics, vol. 10, no. 9, p. 1036, 2021.

[23] K. Markov and T. Matsui, “Music Genre and Emotion Recognition Using Gaussian Processes,” IEEE Access, vol. 2, pp. 688–697, 2014