# EMOTION BASED MUSIC RECOMMENDATION SYSTEM

Meghaa Rhenith R
*Computer Science and Engineering*
*Loyola - ICAM College of Engineering and Technology*
Chennai, India
meghaarhenith.23cs@licet.ac.in

Padmapriya G
*Computer Science and Engineering*
*Loyola - ICAM College of Engineering and Technology*
Chennai, India
padmapriya.23cs@licet.ac.in

Serena J E
*Computer Science and Engineering*
*Loyola - ICAM College of Engineering and Technology*
Chennai, India
serena.23cs@licet.ac.in

Sneha B
*Computer Science and Engineering*
*Loyola - ICAM College of Engineering and Technology*
Chennai, India
sneha.23cs@licet.ac.in

***Abstract*—Songs have long been a preferred form of expression for expressing and comprehending human emotions. The art form that is recognized to connect with a person's emotions the most is music. It has a special capacity to improve one's mood. A user's listing experience will also be enhanced if a recommendation is given depending on his or her mood. The ability to read a person's emotions from their face is crucial. With a camera, the necessary inputs are immediately collected from the human face. This data can then be used to get a list of songs that comply with the mood derived from the input provided earlier. This eliminates the time-consuming and tedious task of manually segregating or grouping songs into different lists and helps in generating an appropriate playlist based on an individual's emotional features. Reliable emotion based classification systems can go a long way in helping us parse their meaning. In this proposed model, we present an effective Emotion Based Music Recommendation System, which recommends music based on the real-time mood of the user. Emotion Detection is done by taking an image of the user's face as input and identifying their mood using convolutional neural networks. Based on the emotion detected, a playlist is recommended to the user which consists of songs suitable to the mood of the user.**

***Index Terms*—Emotion detection, music recommendation, convolutional neural networks, camera.**

## I. Introduction

Songs have long been a preferred form of expression for expressing and comprehending human emotions. The art form that is recognised to connect with a person's emotions the most is music. It has a special capacity to improve one's mood. A user's listing experience will also be enhanced if a recommendation is given depending on his or her mood. The ability to read a person's emotions from their face is crucial. With a camera, the necessary inputs are immediately collected from the human face.

Further analysis, however, demonstrates that listening to music that is completely unrelated to one's mood can have a stressful impact on people. Hence, unwinding with music after work can enhance one's health. A wonderful song should make you feel happy, warm your soul, and boost your spirits. However because there is so much music available, it might be difficult for someone to choose what to listen to from the vast selection. A better platform for all music listeners is provided by an emotion-based music player, which automates song selection and updates playlists on a regular basis based on the user's determined emotion. They help users organise and play tunes in accordance with their moods, reducing stress. So, the objective is to create a recommender system for music and emotions. The analysis of a person's emotional state often involves looking at their face.

Individuals frequently use their facial expressions to convey their feelings. The goal is to create an Emotion Based Music Recommendation System, a web application designed to let users manage extensive playlists with the least amount of effort possible. The suggested model will extract the user's facial expressions in order to ascertain the user's present mood. A playlist of music appropriate for the user's mood will be offered to them once the emotion has been identified.

By playing music that satisfies the user's needs and recording the user's image, the project aids in lifting the user's spirits. We can aid a user in selecting the music they should listen to by creating a recommendation system.

## II. Literature Survey

[1] presents the methodology for an efficient facial expression analysis and classification. Statistical parameters such as entropy, skewness and kurtosis are used to extract the features.180 images of 10 subjects are classified into six categories with 92.2 percent accuracy. SCG training algorithm outperformed in this experiment by classifying the input data in 773 epochs with the average training time of 15 seconds.

[2] In this project the fuzzy relational model for emotion detection is done The degree of a specific human emotion, such as happiness or anger, greatly depends on the degree of MO (Mouth Orbit), EO (Eye Orbit), and the length of EBC. The measurements obtained on MO, EO, and EBC are encoded

into three distinct fuzzy sets: High, Low, and MODERATE. It will be really useful to us as we are making a real-time system which uses Facial expression and recognition.

[3] This project proposes a transfer learning method from face recognition to FER using CNN directly, concatenating both high-level emotion and identity features as Tandem Facial Expression (TFE) features and feeding them to the subsequent fully connected layers. This model concatenates both high-level emotion and identity features as Tandem Facial Expression (TFE) features and feeds it to the subsequent fully connected layers to form a new network.

[4] This project proposes a preference analysis method based on empirical evaluation scores for the selection of general and reliable low level features of music that triggers emotions. It uses an emotion triggering low-level feature selection method based on large number of subjective audience ratings. To check if there is any peculiar competition in the 16 sets of training sets, features are selected from individual training sets with a total of 40 distinct features.

[5] This project aims to compute difference information between the peak expression face and its intra class variation in order to reduce the effect of the facial identity in the feature extraction. A complex CNN network proposed consists of two convolutional layers, each followed by max pooling and four Inception layers. The highest performance approach consists of two parts: spatial image characteristics of the representative expression-state frames are learned using a CNN and temporal changes, making deep learning ill-suited for deployment on mobile platforms with limited resources.

## III. Existing Applications

### A. Music Recommendation System using Content and Collaborative Filtering Methods

The rapid development of mobile devices and the internet has made it possible for us to access different music resources freely. Although the music industry may prefer some genres over others, it is crucial to realize that no human culture on earth has ever existed without music. They developed, put into practice, and evaluated a song recommendation system in this essay. They have made recommendations for songs that users would most enjoy listening to by learning from user listening history and finding relationships between users and songs using the Song Dataset that has been made available. The dataset has over ten thousand songs, and listeners are given recommendations for the finest songs based on the year's top songs by artists, genre, and mood. The top songs and charts of the year are displayed to the listener through an interactive user interface. The dataset gives users the chance to choose their preferred artists and genres when receiving song recommendations.

### B. First Mood-Based Music Recommendation System

A system for matching songs with comparable emotional/emotional content employs vector-distance measurements and an adaptation of the "Valence-Arousal Plane" to make music recommendations. This method seeks to extract the intrinsic features of a music without relying on user data, which makes it fundamentally distinct from collaborative filtering methods. For individuals who don't run a sizable music platform but have billions of pertinent user data points, this is very beneficial. The quick and inexpensive deployment of this strategy, from data collection to implementation of the recommendation algorithm, is another benefit. Last but not least, recommending music from different genres is a desired feature of any real-world recommender system because it helps the user engage with new music and create their own musical tastes.

## IV. Model Training

### A. Convolutional Neural Network

Convolutional Neural Networks(CNN) is a deep neural network, which is most commonly applied to analyzing visual imagery. CNN has numerous applications in image segmentation, object detection, medical image analysis, image captioning, image classification, semantic segmentation, image and video recognition, brain-computer interfaces, financial time series, natural language processing, and recommender systems. CNNs were mainly inspired by the structure of the human brain, in which the biological processes and the connectivity pattern between neurons resembles the organization of the animal visual cortex. CNNs generally consist of Input Layer, Hidden Layer, and Output Layer. The main layers of the CNN are as follows: Convolutional Layer, Pooling Layer, Fully connected layer, Dropout Layer, and Activation Functions. The main purpose of training a model is to find a set of weights and biases that have low loss across all examples.

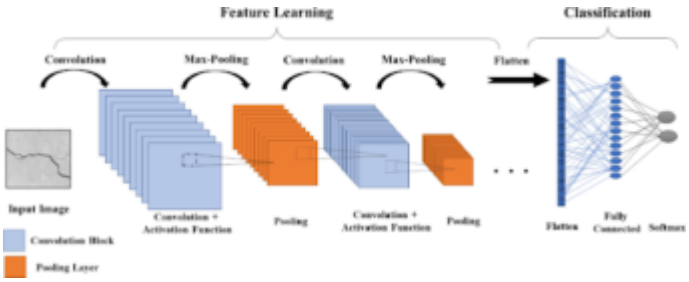


Fig. 1. Convolutional Neural Network Architecture

### B. Sequential Model

The Sequential model is a linear stack of layers. The model training stage consists of model building and compiling using Keras. The main layers of the CNN are Convolutional Layer, Pooling Layer, Fully connected layer, Dropout Layer, and Activation Functions. A Convolutional Layer is the core building block of CNN and is also called the feature extraction layer. It is made from several feature maps and performs a dot product between a matrix, a kernel, and a restricted portion of the receptive field. The kernel slides across the height and width of the image during the forward pass, producing a two-dimensional representation of the image known as an activation map. ReLU activation is used to make all the present

negative values to zero. Conv2D is used as the convolutional layer with different filter sizes ranging from 32 to 128. The pooling layers plays an important role in reducing the computational time and overfitting issues in the CNN architecture. MaxPooling2D is used with pool size 2. On passing a dropout of 0.3, 30 percent of the nodes are dropped out randomly from the neural network. In the proposed system, Dropout is set to 0.25 as anything above resulted in poor performance. Activation functions are used to add non-linearity to the network. It is also called a transfer function. They are generally used to learn and approximate any kind of continuous and complex relationship between the variables of the network. They decide which information of the model should fire in the forward direction and which ones should not at the end of the network. It also defines how the weighted sum of the input can be transformed into an output from the nodes in a layer of the network.

## V. Image Preprocessing

Image Preprocessing is the process of improving the quality of an image by suppressing undesired distortions and enhancing features necessary for the particular application. Images are represented using pixels, which are divided into a grid of pixels and the value of a pixel depends on the type of image. This pixel representation of image matrices can be fed into a machine learning model such as those built using neural networks. Image Preprocessing is performed using the ImageDataGenerator module of Keras.

## VI. Feature Extraction

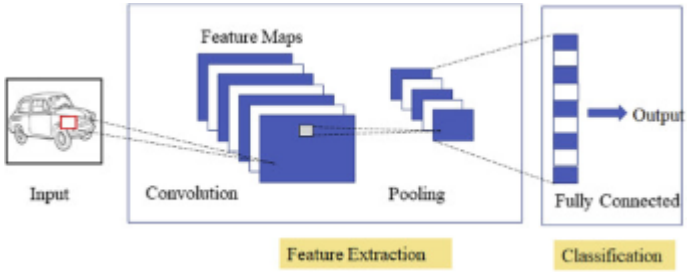


Fig. 2. Feature Extraction and Classification in CNN

Feature extraction techniques are important in image processing and pattern recognition domains. They describe the relevant shape information contained in a pattern and are used to make the task of classifying the pattern easier by a formal procedure. Feature extraction helps to get the best features from segmented images by selecting and combining variables into features, thus reducing the amount of data. The true fact is that CNNs provide automatic feature extraction, which is the primary advantage [6]. In this proposed system, a Convolutional Neural Network (CNN) is used to provide automatic feature extraction, which is the primary advantage. The feature extraction network comprises loads of convolutional and pooling layer pairs, and the pooling layer is used as a dimensionality reduction layer and decides the threshold. During backpropagation, a number of parameters are required to be adjusted, which minimizes the connections within the neural network architecture.

## VII. Proposed System

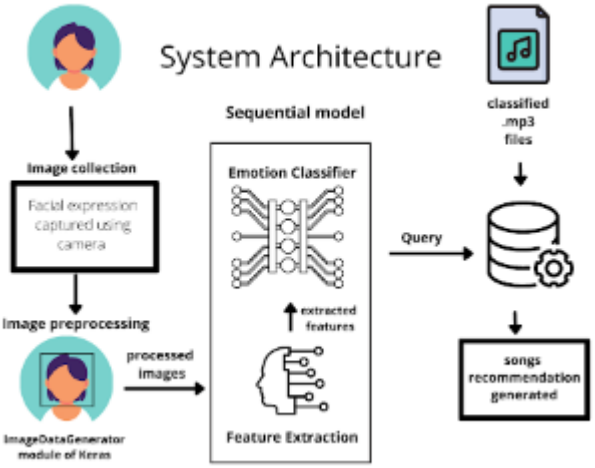


Fig. 3. Architecture Diagram

The model type we have used is sequential which allows us to build a CNN model layer by layer in Keras. Using add function, layers are added to our model. Our model consists of conv2d, maxpooling2d, dropout, dense layers and fallen layer.Conv2d are convolution layers that deals with our input images, which are seen as 2-dimensional matrices. These layers throughout the model have different filter size from 32 to 128 which are the number of nodes in each layer, all with activation function 'relu', Rectified Linear Activation which has been proven to work well in neural networks. Our first layer conv2d takes in input image of shape 48,48,1 with the 1 signifying that the images are greyscale. Maxpooling2d acts as the pooling layer with pool size 2,2. Dropout is set to 0.25 and 0.5 as above which resulted in poor performance. In between the Conv2D layers and the dense layer, we have 'Flatten' layer which serves as a connection between them. Next, to compile our model, we used Adam as an optimizer which adjusts the learning rate throughout training. For loss function, we used 'categorical-crossentropy'. A lower score indicates that the model is performing better. We have used the 'accuracy' metric to see the accuracy score on the validation set when we train the model which 66 percent so further training and fine tuning is required to produce more accurate results.

Emotion Classification is the final step of the proposed system. It is the process of classifying images into different emotion categories based on their features. The dense layer of the sequential model is used as the classifier to classify images based on the output from convolutional layers. Dense is taken as the output layer with 7 nodes, one for each possible outcome that is the emotion detected.Each Layer in the Neural Network contains neurons, which compute the weighted average of its

input and this weighted average is passed through a nonlinear function called ”activation function”. The output has seven neurons with each neuron representing one of the emotions with ”softmax activation” function. Softmax interprets output as probabilities by summing the output to 1. The model will then make its prediction based on the option with the highest probabilityThe app will fetch the playlist of songs from Spotify through spotipy wrapper and recommend the songs by displaying them on the screen.

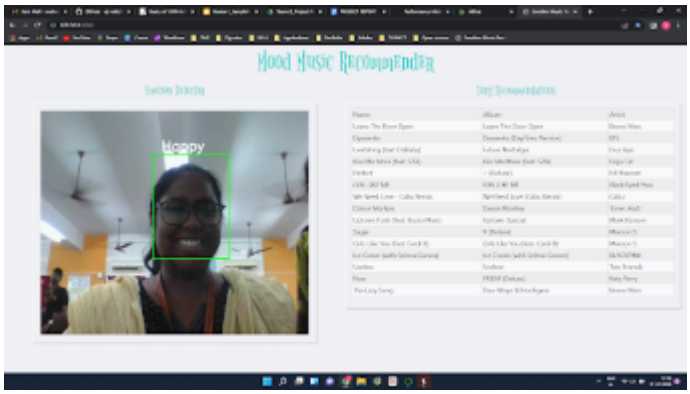


Fig. 4. Emotion detected as “happy” from live webcam

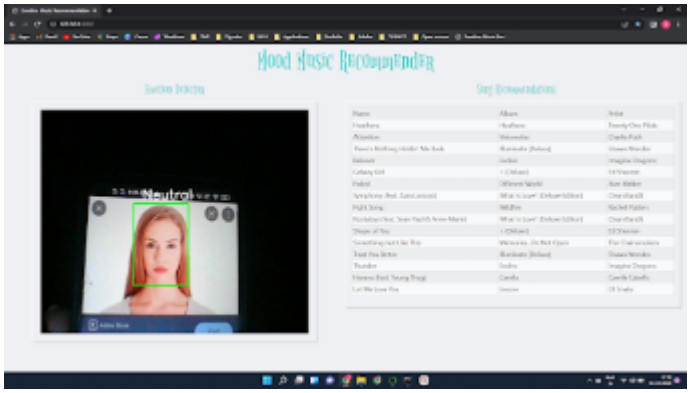


Fig. 5. Emotion detected as “Neutral” from live webcam

## VIII. Results - Accuracy

Performance Metrics are used to measure and summarize the quality of the trained classifier when tested with new data. It is mainly used to evaluate the generalization capability of the trained classifier.

### A. Training Accuracy

Accuracy is a metric that generally describes how the model performs across all classes. It is calculated as the ratio between the numbers of correct predictions to the total number of predictions. Model achieved a training accuracy of 66 percentage.

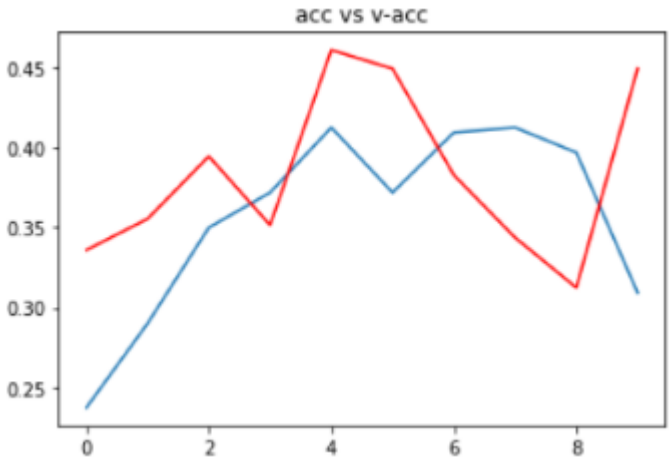


Fig. 6. Graph plotting training accuracy vs validation accuracy

### B. Testing Accuracy

Finally, a test image was input to the system, segmentation and classification was performed on that image and the accuracy was predicted. The image obtained an accuracy of 90 percentage and was classified as sad.

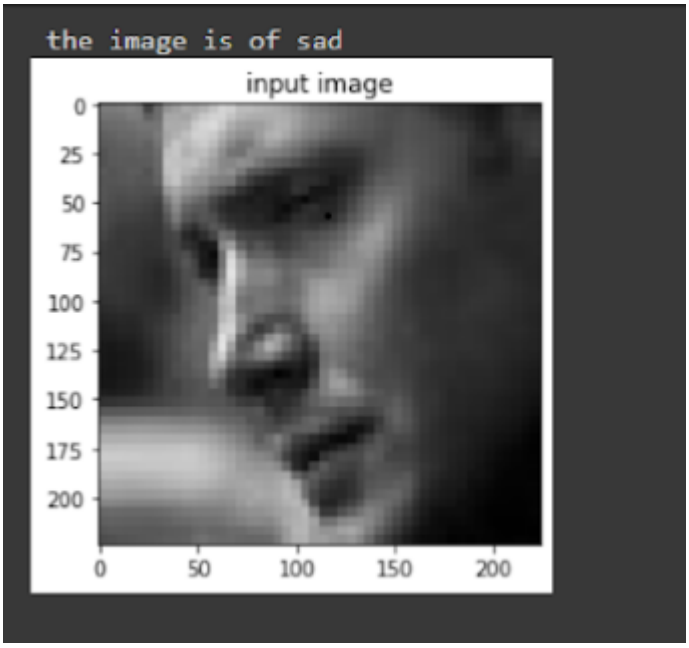


Fig. 7. Predicted Output

## IX. Conclusion

Paying attention to various factors, such as particular context, personal parameters, feelings, and emotions, is highly important to a decision-making process of recommendations. Contemporary music recommendation systems face the gap in personalization, human feelings, contextual preferences and emotional factors while suggesting music. In this model, we proposed an emotion-driven recommendation system with

respect to personalized preferences and particular life and activity contexts. The approach presented in this model is targeted to provide maximum benefits for people from the music-listening experience.

## X. Future Work

Further work on the implementation and testing of the recommendation engine, are considered for the next step when the appropriate amount of the data will be collected. Music creation by artificially intelligent systems with particular music attributes to move states of human emotions can be considered as the further elaboration work in this context. We intend to expand on the concept by adding a heart rate sensor to predict emotion based on heart rate variability(HRV) since sometimes, humans may involuntarily or deliberately conceal their real emotions (so-called social masking). The use of physiological signals can lead to more objective and reliable emotion recognition.

## References


[1] Londhe, Renuka Vrushshen, P Pawar,. (2012). Analysis of Facial Expression and Recognition Based On Statistical Approach. 2.

[2] A. Chakraborty, A. Konar, U. K. Chakraborty and A. Chatterjee, "Emotion Recognition From Facial Expressions and Its Control Using Fuzzy Logic," in IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 39, no. 4, pp. 726-743, July 2009, doi: 10.1109/TSMCA.2009.2014645.

[3] M. Li, H. Xu, X. Huang, Z. Song, X. Liu and X. Li, "Facial Expression Recognition with Identity and Emotion Joint Learning," in IEEE Transactions on Affective Computing, vol. 12, no. 2, pp. 544-550, 1 April-June 2021, doi: 10.1109/TAFFC.2018.2880201.

[4] K. Yoon, J. Lee and M. Kim, "Music recommendation system using emotion triggering low-level features," in IEEE Transactions on Consumer Electronics, vol. 58, no. 2, pp. 612-618, May 2012, doi: 10.1109/TCE.2012.6227467.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.Ko BC. A Brief Review of Facial Emotion Recognition Based on Visual Information. Sensors (Basel). 2018 Jan 30;18(2):401. doi: 10.3390/s18020401. PMID: 29385749; PMCID: PMC5856145

[6] R. Miotto, F. Wang, S. Wang, X. Jiang, J.T. Dudley, Deep learning for healthcare: review, opportunities and challenges, Briefings in Bioinformatics (2017)