Human Visuals for Vibrant Keytunes – HVVK an Emotion Detection Music Recommendation System using Spotify API

Vania Panjwani[1], Harsha Gotmare[2], Vedant Masane[3] , Prof. Dr Rashmi Jaiswal[4]

[1]Student,Dept.of Comp Sci & engineering, COETA, Akola, Maharashtra, India
[2] Student, Dept. of Comp Sci & engineering, COETA, Akola, Maharashtra, India
[3] Student, Dept. of Comp Sci &engineering, COETA, Akola, Maharashtra, India
[4] Assistant Professor, Dept. of Comp Sci & engineering, COETA, Akola, Maharashtra, India

---***---

Abstract - *The swift progress in mobile and internet technology has granted us unrestricted access to an extensive array of music resources. In the music industry, certain musical genres could be more well-liked than others. At the moment, users of music-listening apps have to make special updates to the static playlists to reflect their own preferences. Music recommendation systems have become an integral part of music listening. However, most traditional recommendation systems rely on user activity data or metadata, which is not enough to capture the emotional resonance of music. Given that music has a strong emotional impact on listeners, personalized music recommendation systems that include users' emotional states are highly desired.*

Two different models and methods were available for this purpose: FER (Facial Emotion Recognition), which detects mood from facial expressions, and Music Classification Models, which select songs. To provide the user with recommendations that are accurate and efficient, we are trying to integrate these two systems. To recommend music that is appropriate for the user's emotions, we will be developing a system in this project that will allow us to collect the user's real-time emotions through conversation or other methods.

***Key Words*: Recommendation System, Facial Emotion Recognition, Interactive UI, Mood-based music classifier, Spotify API.**

1. INTRODUCTION

Music now plays a vital role in our lives in the digital age by providing a soundtrack for our feelings and experiences. With the introduction of streaming services, we now have an incredible amount of music at our fingertips. But finding music that speaks to our current emotional states and connecting with them is the difficult part—not the availability of music. In an attempt to create customized playlists that suit a listener's tastes and mood, music recommendation algorithms have surfaced as a solution to this problem.

The main goal of this project is to close the gap that exists between a user's emotional state and their experience listening to music. As everyone knows, our feelings have a big impact on the kind of music we listen to. For example, someone who is happy could enjoy lively, uplifting music, but someone who is depressed might prefer more reflective or peaceful sounds. Conventional music recommendation systems could miss the user's current emotional state because they frequently rely on the user's past and preferences. Our approach uses AI, Spotify's vast music catalog, and facial emotion recognition to overcome this constraint.

2. PROBLEM STATEMENT

Even though music evolved a lot, streaming services have also seen new developments, but still users face the challenge of not discovering music that matches their mode and emotion. Traditional music recommendation was based on physical input from the user or from the previous metadata that provided recommendations to the player and then to the user. Music has a strong emotional impact on the listeners, but the traditional system fails to capture the emotions of the listener. Therefore, the aim of this paper is to create a system that uses machine learning algorithms to present a cross-platform music player that makes recommendations for music depending on the user's current mood as seen through a webcam.

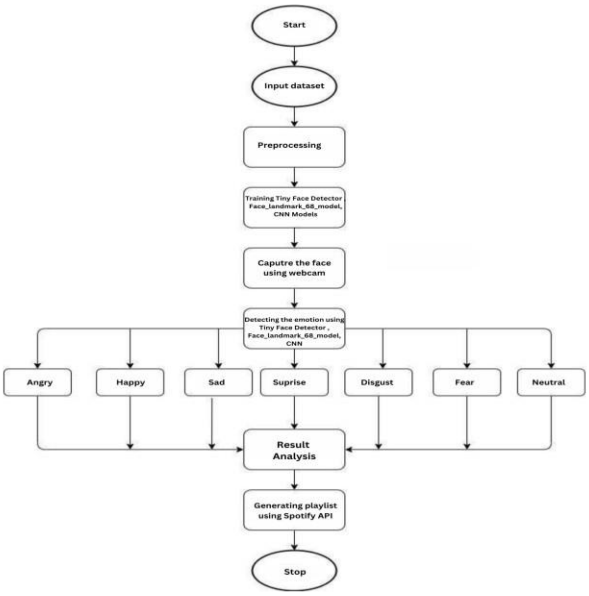
3. LITERATURE REVIEW

Facial expressions convey the person's present mental state. When expressing feelings to others, we typically do it using nonverbal cues including tone of voice, facial expressions, and hand gestures. Preema et al. [1] claimed that making and maintaining a big playlist takes a lot of effort and time. According to the publication, the music player chooses a tune based on the user's present mood. Playlists based on mood are created by the application by scanning and categorizing audio files based on audio attributes. The Viola-Jonas method, which is used for face detection and facial emotion extraction, is utilized by the program. The categorization process employed Support Vector Machine (SVM) to extract data into five main universal emotions, such as anger, joy, surprise, sad, and disgust.

Ayush Guidel et al. claimed in [2] that facial expressions are a simple way to read a person's mental state and present emotional mood. Basic emotions (happy, sad, angry, excited, surprised, disgusted, fearful, and neutral) were taken into consideration when developing this system. In this research, face detection was implemented by convolutional neural network. Stories about music being a "language of emotions" are common throughout the world.

4. METHODOLOGY

4.1 SOFTWARE ARCHITECTURE


Start
Input dataset
Preprocessing
Training Tiny Face Detector , Face_landmark_68_model, CNN Models
Caputre the face using webcam
Detecting the emotion using Tiny Face Detector , Face_landmark_68_model, CNN
Angry
Happy
Sad
Suprise
Disgust
Fear
Neutral
Result Analysis
Generating playlist using Spotify API
Stop

Flow Chart- 4.1: SOFTWARE ARCHITECTURE

4.2 Data Gathering

Data gathering is the crucial step as it provides the important data upon which training, testing, and validating machine learning models occurs , data gathering is an essential step in the machine learning process. The amount and quality of data collected is a critical factor in deciding how accurate, reliable, and broadly applicable machine learning models may be. The dataset's characteristics, like its complexity, diversity, and quantity of features, affect the project's efficiency and accuracy. Results will be more effective the more diversified the dataset is. In conclusion, data collection is critical to machine learning since it directly impacts the models' resilience, accuracy, and capacity for generalization. To make sure that the models can be applied to new data and can be broadly generalized, it is essential to collect a variety of representative data sets.

FER-2013:

The FER-2013 dataset was generated by compiling the outcomes of a Google image search for each emotion together with its synonyms. 35,887 48x48 pixel grayscale photos make up the dataset. One of seven feeling categories—angry, disgusted, afraid, glad, sad, surprised, or neutral—is assigned to each image. Crowd-sourcing was used to obtain the classifications, asking human annotators to categorize the emotions depicted in the photos. The FER2013 database is divided into training and testing datasets.28,709 photos make up the training set, and 3,589 images make up the test set.

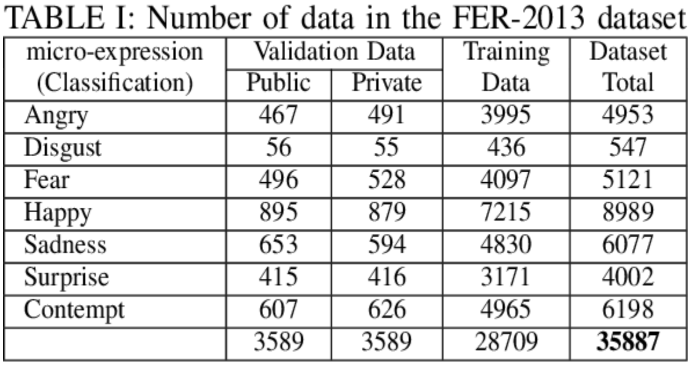


TABLE I: Number of data in the FER-2013 dataset

micro-expression	Validation Data		Training	Dataset
(Classification)	Public	Private	Data	Total
Angry	467	491	3995	4953
Disgust	56	55	436	547
Fear	496	528	4097	5121
Happy	895	879	7215	8989
Sadness	653	594	4830	6077
Surprise	415	416	3171	4002
Contempt	607	626	4965	6198
	3589	3589	28709	**35887**

Fig-1: Fer Dataset

4.3 Model Selection

Tiny Face Detector:

The term "Tiny Face Detector" describes a particular kind of object identification model intended to identify faces in pictures or videos. Its compact size, as the name implies, makes it ideal for applications like embedded systems and mobile devices where computational resources are scarce. The designation "Tiny" usually denotes that the model has undergone efficiency optimization, frequently using methods like as quantization and model compression.

A deep learning model created especially for the purpose of identifying faces in photos is called the Tiny Face Detector model. It is a variation on the popular YOLO (You Only Look Once) object detection architecture, which excels at tasks requiring quick and accurate object detection in real time.

We process the entire image using a single neural network. The image is divided into regions by this network, which then forecasts bounding boxes and probabilities for each region. The expected probabilities are used to weight these bounding boxes. Since it examines the entire image during testing, the global context of the image informs its predictions.

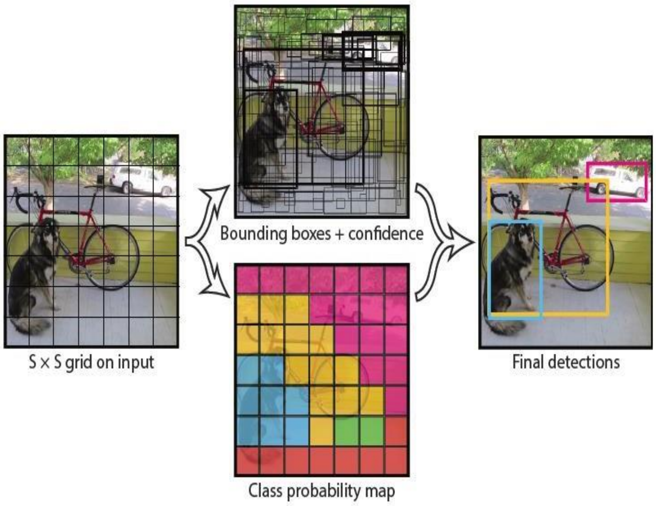


Fig-2 :Tiny face detection

Face_landmark_68_model:

The term "face_landmark_68_model" describes a particular kind of facial landmark detection model that can identify 68 important locations on an individual's face. In computer vision and image processing applications, facial landmark detection is essential, especially for face analysis, facial emotion identification, facial alignment, and augmented reality applications.

Typical landmarks identified by a "face_landmark_68_model" consist of:

- The areas surrounding the eyes, such as the corners,

eyelids, and iris centers.

- The inner and outer corners of the eyebrow points.
- The nose's surrounding points (tip, nostrils, bridge).
- The points on either side of the mouth (lips, corners).
- Points on the chin and jawline.

A lot of machine learning approaches, especially deep learning architectures like convolutional neural networks (CNNs), are used to develop facial landmark identification models. These models have been developed using

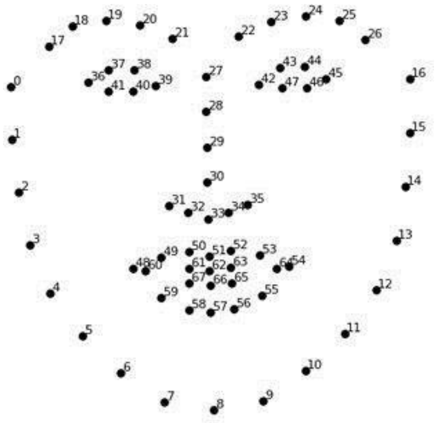


Fig-3: Face_landmark_68_model

Convolutional Neural Network (CNN):

CNNs are the foundation of our model architecture because they work effectively for image-based tasks like facial expression recognition. Learning spatial hierarchies in visual data is a skill that CNNs excel at. The acronym CNN represents Convolutional Neural Network. This kind of artificial neural network (ANN) is especially useful for tasks involving picture recognition and classification and is frequently employed in deep learning. CNNs are very good at processing spatial input, such as images, because they are modeled after the anatomy and physiology of the animal visual cortex.

CNNs' fundamental units are convolutional layers. They employ learnable filters or kernels to apply convolution operations to the input data. Local patterns and characteristics are extracted from the input image by swiping these filters over it. Different elements of the input, such as edges, textures, or forms, are captured by each filter. In order to reduce the spatial dimensions of the feature maps produced by the convolutional layers while preserving significant features, pooling layers down sample the feature maps. CNNs are able to identify intricate patterns and relationships in the data by introducing non-linearity into the network using activation functions. CNNs frequently employ the sigmoid, tanh, and ReLU (Rectified Linear Unit) activation functions.

Traditional neural network layers, referred to as fully linked or dense layers, have connections between every neuron in the layer above and below it. Towards the end of the network, these layers are usually utilized to map extracted features to certain output classes. Small matrices called filters or kernels are applied to the input image to carry out convolution processes. These filters extract elements like edges, textures, and patterns by swiping over the input.

The learnt features of the input image are represented by feature maps, which are the result of the convolutional layers. Every feature map captures distinct facets of the input and is matched to a particular filter. Supervised learning is the method used to train CNNs to map input images to the appropriate labels or classes. In order to reduce the discrepancy between expected and actual outputs, the network uses optimization techniques like gradient descent to modify its parameters (weights and biases) during training.

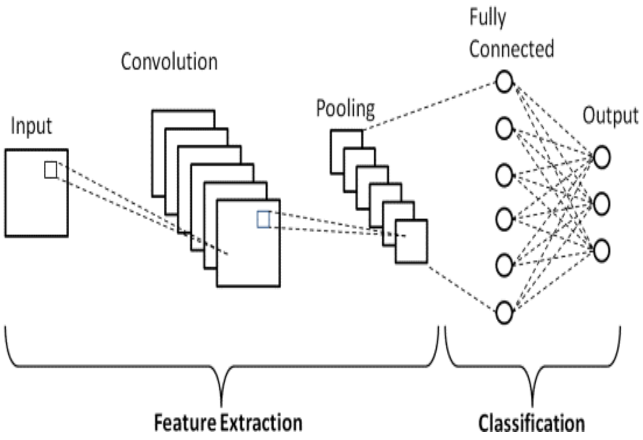


Fig-4: Convolutional Neural Network (CNN)

4.4 Model Building

As the actual development and use of theoretically proposed models and approaches are performed in this phase, model building is an important step in the machine learning process. The model that we suggested or selected in the earlier stages is trained on the relevant datasets and then applied to provide predictions for the fresh data.

Before Actual development of models there are some prerequisite step in machine learning as data preprocessing .

Data preprocessing is a step in which the raw data is transformed into a data format which can be easily analyzed by machine learning algorithm. The data preprocessing step involves cleaning, normalize and transformation of raw data to use it for training and testing machine models.

The FER model's evident input is an image, which can be found in the dataset in a variety of file types, including.PNG and others. Therefore, in order to make it easier for the model to assess, we transformed those image formats into an array. An additional method for preparing data on the FER-2013 dataset is to scale the data to fit inside a certain range or distribution. Thus, this may aid in enhancing our algorithm's performance.

FER model

As we start execution the user provides the input , the input is processed frame by frame. As the finalized models for facial recognition are Tiny face detector , Face Landmark 68 _Model and Convolutional Neural Network (CNN), we will be using the three model for different part of there specialties. The first model will be the Tiny face detector after which we will use Face Landmark 68 _Model and at last the Convolutional Neural Network (CNN).

As the process starts the input image through the webcam is taken the tiny face detector divides the frame into smaller frames. Frames having the same element or connected elements are grouped together grouped frames are assigned a color and colored processed frames are filtered to remove low confidence detections and show results of high confidence.

This high confidence output is taken as input for both the Face Landmark 68 _Model and the Convolutional Neural Network (CNN) model. If we first consider Face Landmark 68 _Model as the name says

68 landmarks are pointed onto the face and the landmark points are matched with the FER2013 dataset data to again get the output of the highest expression. Second, we consider a Convolutional Neural Network to mark the neural network layer on the face. After this pooling layer is used to map features. Then high-level reasoning is done on the features extracted. The number of nodes in the output layer depends on the number of classes in the classification task. The classification task is the emotional classifications present in the FER2013 database. At the end outputs of both the models are classified again with the dataset to get the highest expression by taking both models into consideration.

First, the necessary datasets and libraries are loaded before the model is really trained. Subsequently, we examined the real mood configuration, or the quantity of photos included in each mood group. Following examination, it becomes clear that there are various mood-related images in distinct ratio. Since the photos in the FER-2013 dataset are 48 by 48 in size, we choose to downsize our input image to a size that is comparable in the hopes of obtaining an accurate prediction.

We divided our dataset into 80% training and the remaining 20% testing portions for training purposes, making sure that the division was entirely random. Transfer learning is a deep learning technique in which a previously learned model is applied to a new task. Utilizing transfer learning enables us to take advantage of the information and learnt representations of an existing model rather than creating and training a new one from scratch. All the models. Tiny face detector , Face Landmark 68 _Model and Convolutional Neural Network (CNN) support transfer learning.

4.4 Music recommendation

For music recommendation we are using dynamic data base through spotify API. The output of model generation are taken as input for music recommendation . The emotion value obtained such as value in FER dataset as angry, disgust, fear, happy ,etc is passed on to the API. The api then provides back with the songs corresponding to the emotions . The provided is output is again classified as Albums , Artist and Playlist. The user has the freedom to choose the song from the given three classifications.

Spotify integration satisfies all the expectations of the user. Also it provides with a large selection of songs to enjoy in very mood. The Spotify API offers developers a number of extra features that might aid in the creation of top-notch projects. Managing personal libraries, searching for desired information, retrieving data from targeted albums, songs, or artists, and control and interaction are the major features of these functions. Furthermore, it can offer us a range of songs in addition to the audio elements of every song. These audio features are essentially the same as our dataset's acoustic features. Spotify, the official Spotify collection, addresses these characteristics of the tracks.

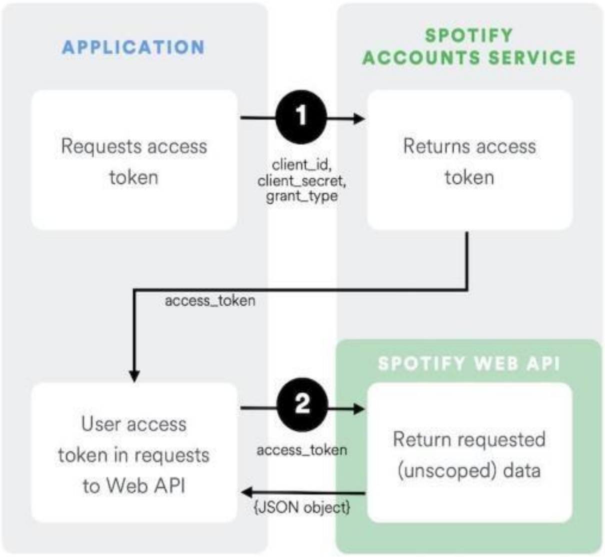
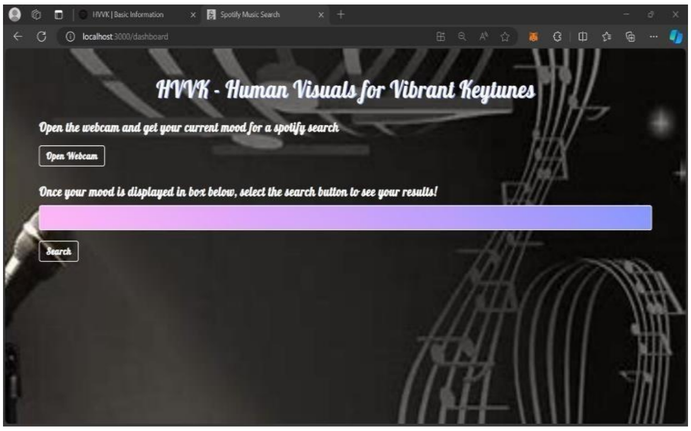


Fig-5: Spotify API integration using Developer Account

5. RESULT

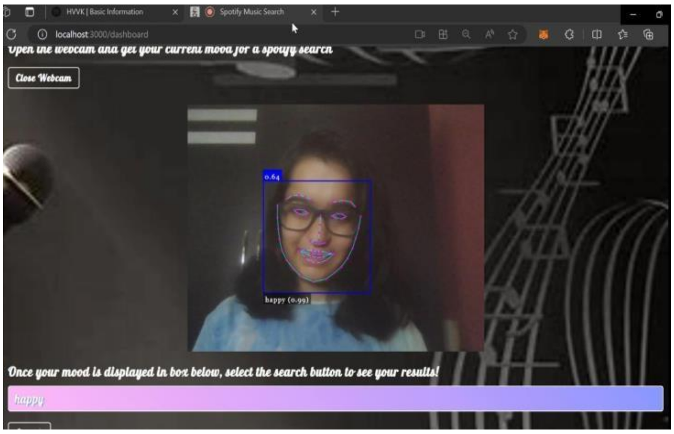


HVVK - Human Visuals for Vibrant Keytunes

Open the webcam and get your current mood for a spotify search

Once your mood is displayed in box below, select the search button to see your results!

Fig-6: the main screen of project will ask to Open the Webcam



Open the webcam and get your current mood for a spotify search

Once your mood is displayed in box below, select the search button to see your results!

Fig-7: The Webcam will detect your real time emotion and display the same

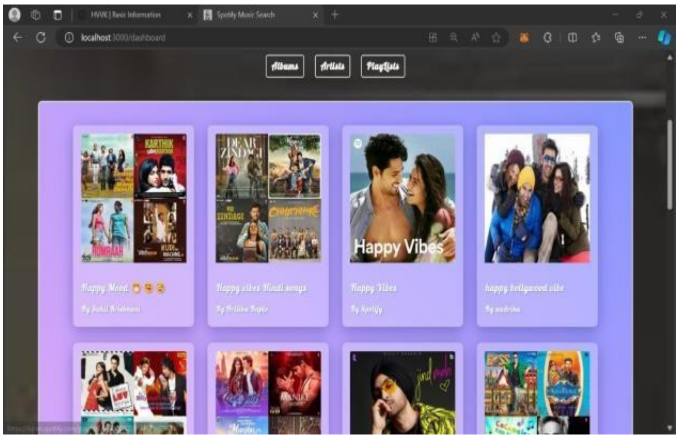
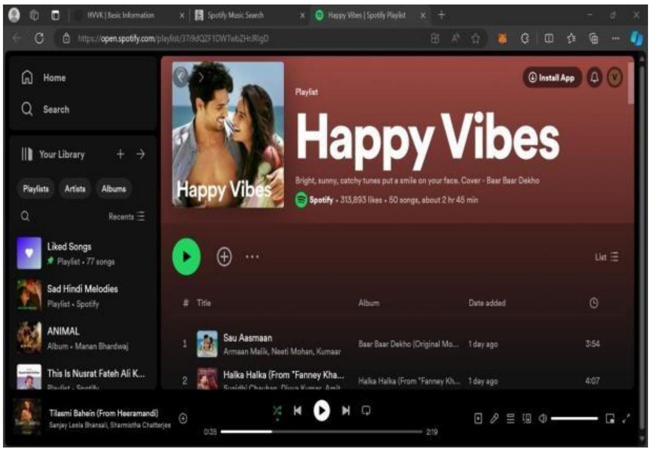


Fig-8: Click the Search Button, the Albums of the emotion detected would be displayed first



Happy Vibes

Bright, sunny, catchy tunes put a smile on your face. Cover - Baar Baar Dekho

Spotify • 313,893 likes • 50 songs, about 2 hr 45 min

#	Title	Album	Date added	
1	Sau Aasmaan Armaan Malik, Neeti Mohan, Kumaar	Baar Baar Dekho (Original Mo...	1 day ago	3:54
2	Halka Halka (From "Fanney Kha...	Halka Halka (From "Fanney Kh...	1 day ago	4:07

Fig-9: Click the Search Button, the Albums of the emotion detected will be displayed first

6. FUTURE SCOPE

The use of numerous novel concepts in CNN's architecture has shifted project priorities, particularly in the field of computer vision. To study innovations in CNN's architecture is an encouraging study area, and has the potential to become one of the most utilized AI techniques. Ensemble learning is an upcoming project area in CNN. By extracting distinct semantic representations, the model can improve the generalization and resilience of many categories of images by combining multiple and diverse designs. In picture segmentation tasks, although it performs well, a CNN's ability as a "generative learner" is limited.

The use of CNNs' generative learning capabilities throughout feature extraction phases can improve the model's representational power. At the intermediate phases of CNN, fresh examples can be incorporated to improve the learning capability by using auxiliary learners .Attention is a crucial process in the human visual system for acquiring information from images. Furthermore, the attention mechanism collects the crucial information from the image and stores its context in relation to the other visual components. In the future, the project could be conducted to preserve the spatial importance of objects and their distinguishing characteristics during subsequent stages of learning.

7. CONCLUSION

Our project will be a cutting-edge trending technology utilization, as the benefits and power of AI-powered apps are becoming more and more popular. We give a summary of how music might influence a user's mood in this system, along with guidance on selecting the appropriate tunes to lift users' spirits. Emotions of the user can be detected by the developed system. The system was able to identify the following emotions: neutral, shocked, furious, sad, and pleased. The suggested algorithm identified the user's sentiment and then showed them a playlist with songs that matched their mood. Processing a large dataset requires a lot of RAM in addition to CPU power. Development will become more appealing and hard as a result. The goal is to develop this application as cheaply as feasible while using a standardized platform. Our face emotion recognition-based music recommendation system will lessen users' work in making and maintaining playlists.

In the future, we hope to create a more precise model as well as a modem that can reliably identify all six main emotions: happy, sad, angry, neutral, fear, disgust, and surprise. We hope to create an easily installable Android app in the future that will provide an emotion-based music recommendation system for our phones. Additionally, we intend to create unique features that would offer quotes from notable figures based on the user's emotions. For example, if the user is identified as down, a song recommendation and a quote from the song will be shown to uplift the user. Individuals often listen to songs while doing other things. If a quote inspires them or makes them feel engaged, they are more likely to post it on social media, which helps inspire at least a small number of people. Another feature is that points may be earned to buy singer albums; these points are accumulated based on how frequently a user uses the application.

REFERENCES

[1] Preema J.S, Rajashree, Sahana M, Savitri H, Review on facial expression-based music player, International Journal of Engineering Research & Technology (IJERT), ISSN-2278-0181, Volume 6, Issue 15, 2018.

[2] AYUSH Guidel, Birat Sapkota, Krishna Sapkota, Music recommendation by facial analysis, February 17, 2020

[3] Ramya Ramanathan, Radha Kumaran, Ram Rohan R, Rajat Gupta, and Vishalakshi Prabhu, an intelligent music player based on emo-tion recognition, 2nd IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions 2017. https://doi.org/10.1109/CSITSS.2017.8447743

[4] Shlok Gilda, Husain Zafar, Chintan Soni, Kshitija Waghurdekar, Smart music player integrating facial emotion recognition and music mood recommendation, Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India, (IEEE),2017. https://doi.org/10.1109/WiSPNET.2017.8299738

[5] CH. sadhvika, Gutta.Abigna, P. Srinivas Reddy, Emotion-based music recommendation system, Sreenidhi Institute of Science and Technology, Yamnampet, Hyderabad; International Journal of Emerging Technologies and Innovative Research (JETIR) Volume 7, Issue 4, April 2020.

[6] Madhuri Athvale, Deepali Mudale, Upasana Shrivatsav, Megha Gupta, Music Recommendation based on Face Emotion Recognition, Department of Computer Engineering, NHITM, Thane, India

[7] Sheela Kathavate, Music Recommendation System using Content and Collaborative Filtering Methods, Department of Information Science and Engineering BMS Institute of Technology and Management Bangalore, India

[8] Research Prediction Competition, Challenges in representation learning: facial expression recognition challenges, Learn facial expression from an image, (KAGGLE).

[9] Emotion Recognition on FER-2013 Face Images Using Fine-Tuned VGG-16. Gede Putra Kusuma, Jonathan, Andreas Pangestu Lim.