EmoTune: Song Recommendation through Facial Emotion Recognition

Monica Gullapalli[1] and Pradyumna Chippigiri[2]

University Of Colorado Boulder

1 Abstract

We propose a radical shift in music recommendation at "EmoTune," introducing real-time facial emotion recognition as a critical factor. While leaving the existing systems do not consider the listener's emotional state during the music selection process, our system relies on a personally customized deep convolutional neural network trained on the FER 2013 dataset which validates music recommendations against the user's mood. Through comparative experiments with well-known models such as VGG16, we have shown significant superior performance in terms of accuracy and user satisfaction. Moreover to make this application more tailored to the user, we made it a multi-modal application where even the song playlist generation is performed by understanding the acoustic features in tunes and understanding their impact on users for which we trained our music dataset over different algorithms and confirmed SVM Model to be of best use. Additionally, our results have exhibited the consistency of EmoTune across different dataset sizes, leading to crucial implications in applicable life scenarios for emotionally-centered music listening.

2 Introduction

The human face serves as a vital indicator of a person's emotional state, making it an invaluable resource for understanding one's mood, and music profoundly influences human emotions, serving as a reflection of our feelings or as a tool to change our moods. The aim of this project is to create a system whose ability is to discern and respond to our emotions through facial analysis. The goal of the project is to create a listening experience that is not just tailored to our musical tastes but synchronized with our emotional rhythms thus enhancing the listening experience and thus providing therapeutic benefits by aligning music to emotional needs. Thereby, enhancing user engagement through personalized emotion-driven recommendations. This could open new pathways for mental wellness, where music becomes a tool for emotional regulation and therapy. The motivation behind our work is to underscore the growing necessity to integrate emotional intelligence into technological solutions, particularly in domains such as human-computer interaction, virtual reality, and affective computing, where understanding and responding to human emotions can significantly enhance user engagement and satisfaction.

Current music recommendation platforms rely heavily on algorithms based on user history and collaborative filtering. This method overlooks the listener's immediate emotional state, resulting in suggestions that might fit musical tastes but not current moods. This gap highlights a crucial oversight: the lack of real-time emotional understanding in music curation, limiting the potential for music as a mood regulation tool.

In this project, we introduce an innovative solution by integrating facial emotion recognition with music recommendations. This approach allows for dynamic, emotion-driven music suggestions that align with the listener's current state of mind. By addressing the immediate emotional context, "EmoTune" promises a more engaging and satisfying listening experience, setting a new standard for personalized music service.

3 Related Work

Topic: Music Recommendation system
Reference 1: Singh et al. [1]
Reference 2: Durga et. al. [2]
Reference 3: Chen et. al. [3]
Reference 4: Sakti et. al. [4]

Our work is different from these as Singh et al. [1] paper does not integrate real-time facial expression detection with music recommendation for personalization. The above 4 papers leverage both user data and song data to recommend songs and these models integrate collaborative filtering, content-based filtering, and other data processing methods to create a more robust and versatile recommendation but don't make use of the current real-time emotion of the user.

Topic: Emotion Detection System
Reference 1: Amara et al. [5]
Reference 2: Pandey et. al. [6]
Reference 3: Chandraprabha et. al. [7]

Our work is different from these as the above 3 papers discuss emotion detection based on facial expressions using pre-trained models such as VGG-16 and VGG-19. But we will be creating a custom DCNN(Deep Convulutional Neural network) and will be comparing the results with the existing models. Also our real-time detection system could detect nuanced facial expressions along with rapid expression change detection.

4 Methods

The approach we have to cater to this cause and solve the problem has a multimodal solution. The entire application is divided into 2 segments, the first segment is such does the Music Recommendation based on the mood and the second does the Emotion detection using Facial recognition.

The first system works on training the song data and predicts the song mood based on its valence, danceability and energy. Then Spotify API is used to create playlists based on each mood. These playlists are then used to display to the user based on their facial emotion.

The second system works based on training a model with image input data, which is followed by the recommendation of a Spotify playlist based on the user's face being captured by the application's front camera to determine what emotion the user is experiencing.

The core idea is to use a dataset of human photos representing various emotions to feed into a TensorFlow machine learning model in Python. Once the model is loaded, you can use cv2 to capture videos that aid in understanding the user's sentiment. This will be followed by the user getting redirected to the playlist that has been curated based on their mood. The application aims to enhance or improve the mood of a user while they listen to music.

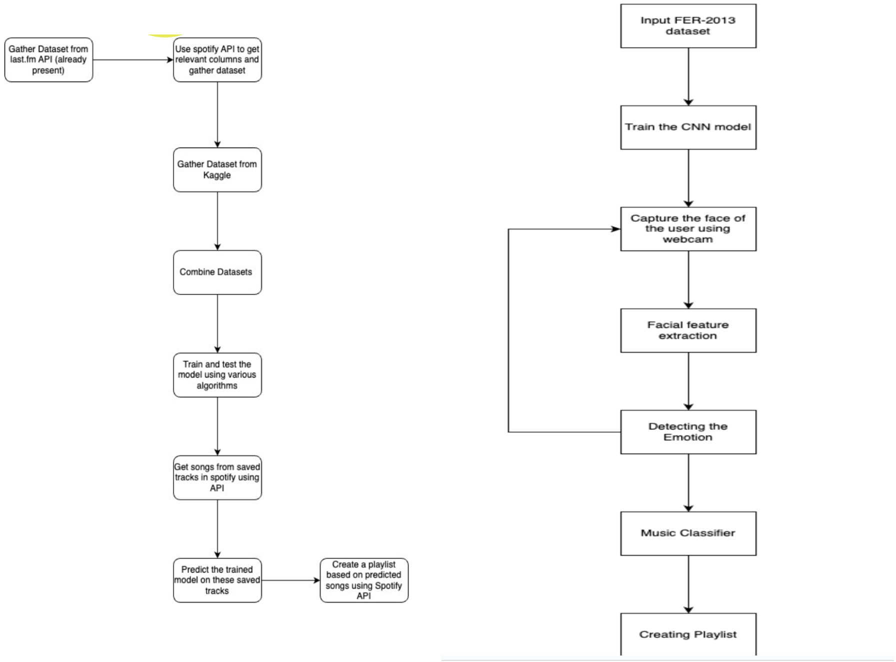

Gather Dataset from last.fm API (already present)
Use spotify API to get relevant columns and gather dataset
Gather Dataset from Kaggle
Combine Datasets
Train and test the model using various algorithms
Get songs from saved tracks in spotify using API
Predict the trained model on these saved tracks
Create a playlist based on predicted songs using Spotify API
Input FER-2013 dataset
Train the CNN model
Capture the face of the user using webcam
Facial feature extraction
Detecting the Emotion
Music Classifier
Creating Playlist

Fig. 1. The EmoTune flask application

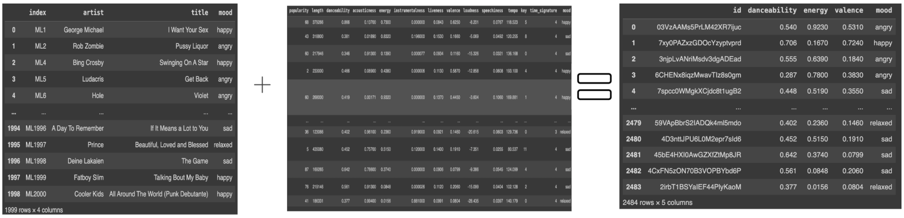
4.1 System Overview

The system will be able to detect the user's emotion in real-time using the device camera. The system will then classify the user emotion into categories like : Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise followed by user redirection to playlists that will suit their mood.

4.2 Dataset

Datasets for Song Recommendation The dataset, sourced from last.fm comprises user-generated tags and acoustic features of music tracks. The dataset includes various attributes of songs that are integral to understanding their impact on listeners' emotions. The second dataset was obtained from a Kaggle competition, which focuses specifically on the mood associated with different music tracks. It features similar acoustic attributes to those found in the last.fm dataset, along with mood labels that categorize tracks into emotional states such as happy, sad, relaxed, angry, disgusted, fear, surprised.

To leverage the strengths of both datasets, we combined them into a unified dataset. The merging process involved aligning tracks by their unique identifiers (song IDs) and then consolidating the relevant columns. The resultant combined dataset includes the following key features: 'song id', 'danceability', 'energy', 'valence', and 'mood'. This dataset serves as the foundation for training our music recommendation models, allowing us to generate playlists that are not only tailored to specific musical tastes but also aligned with the listener's current emotional state.



	index	artist	title	mood
0	ML1	George Michael	I Want Your Sex	happy
1	ML2	Rob Zombie	Pussy Liquor	angry
2	ML4	Bing Crosby	Swinging On A Star	happy
3	ML5	Ludacris	Get Back	angry
4	ML6	Hole	Violet	angry
...	...	...	...	...
1994	ML1996	A Day To Remember	If It Means a Lot to You	sad
1995	ML1997	Prince	Beautiful, Loved and Blessed	relaxed
1996	ML1998	Deine Lakaien	The Game	sad
1997	ML1999	Fatboy Slim	Talking Bout My Baby	happy
1998	ML2000	Cooler Kids	All Around The World (Punk Debutante)	happy

1999 rows × 4 columns

popularity	length	danceability	acousticness	energy	instrumentalness	liveness	valence	loudness	speechiness	tempo	key	time_signature	mood
68	379266	0.866	0.13700	0.7300	0.000000	0.0843	0.6250	-8.201	0.0767	118.523	5	4	happy
43	318800	0.381	0.01890	0.8320	0.196000	0.1530	0.1660	-5.069	0.0492	120.255	8	4	sad
60	217946	0.346	0.91300	0.1390	0.000077	0.0934	0.1160	-15.326	0.0321	136.168	0	4	sad
2	233000	0.466	0.08900	0.4380	0.000006	0.1130	0.5870	-12.858	0.0608	193.100	4	4	happy
60	268000	0.419	0.00171	0.9320	0.000000	0.1370	0.4450	-3.604	0.1060	169.881	1	4	happy
...	...	...	...	...	...	...	...	...	...	...	...	...	...
36	123066	0.402	0.96100	0.2360	0.919000	0.0921	0.1460	-20.615	0.0603	129.736	0	3	relaxed
5	435080	0.452	0.75700	0.5150	0.120000	0.1400	0.1910	-7.351	0.0255	80.537	11	4	sad
87	169265	0.642	0.78600	0.3740	0.000000	0.0906	0.0799	-9.386	0.0545	124.099	4	4	sad
76	219146	0.561	0.91300	0.0848	0.000026	0.1120	0.2060	-15.099	0.0404	102.128	2	4	sad
41	186331	0.377	0.99400	0.0156	0.881000	0.0991	0.0804	-28.435	0.0397	140.179	0	4	relaxed

	id	danceability	energy	valence	mood
0	03VzAAMs5PrLM42XR7ijuc	0.540	0.9230	0.5310	angry
1	7xy0PAZxzGDOcYzyptvprd	0.706	0.1670	0.7240	happy
2	3njpLvANriMsdv3dgADEad	0.555	0.6390	0.1840	angry
3	6CHENx8iqzMwavTIz8s0gm	0.287	0.7800	0.3830	angry
4	7spcc0WMgkXCjdc8t1ugB2	0.448	0.5190	0.3550	sad
...	...	...	...	...	...
2479	59VApBbrS2IADQk4ml5mdo	0.402	0.2360	0.1460	relaxed
2480	4D3nttJPU6L0M2epr7sld6	0.452	0.5150	0.1910	sad
2481	45bE4HXI0AwGZXfZtMp8JR	0.642	0.3740	0.0799	sad
2482	4CxFN5zON70B3VOPBYbd6P	0.561	0.0848	0.2060	sad
2483	2irbT1BSYaIEF44PlyKaoM	0.377	0.0156	0.0804	relaxed

2484 rows × 5 columns

Dataset for Emotion Recognition The dataset for the Emotion detection using facial emotion recognition application is FER 2013 which has the emotion detection dataset split into training and validation datasets, which has 7 different features based on the labels attached to them like Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise. In split of the dataset there is a total of 28709 Images and 3589 Images in the training and the Validation split respectively. The dataset is utilised enitrely to train and validate the image data as input, the test set was directly gonna be the users when they use the application, from the camera available on their device.

4.3 Algorithms Used

1. **Music Recommendation System:** In our EmoTune system, we employ a suite of machine-learning models to enhance our mood-based song recommendation capabilities. The selection includes a **Support Vector Machine**

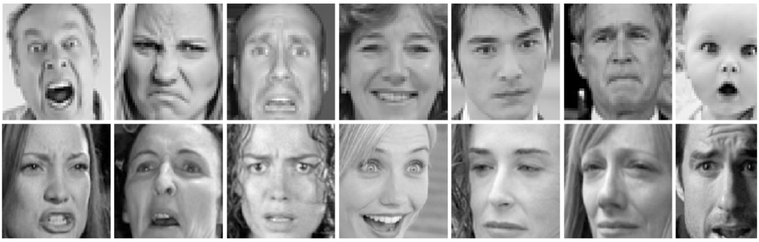


Fig. 2.

(SVM), which is adept at handling high-dimensional spaces and uses kernel methods to manage non-linear data, making it highly effective for classifying songs into different emotional categories. A **Decision Tree** model offers straightforward decision paths that easily categorize songs by their acoustic features into specified moods, providing clear interpretability. **Logistic Regression**, adapted for multiclass challenges, estimates the probability of each mood, giving a baseline for algorithmic performance. Lastly, the **K-Nearest Neighbors (KNN) Classifier** uses the similarity of songs to dynamically adjust recommendations based on user feedback, continuously refining the personalization of the playlist. Together, these models form a robust framework for our song recommendation system, ensuring that the music suggestions not only match the listener's preferred genres but also align seamlessly with their current emotional state.

2. **Emotion Detection System Using Facial Recognition:** The main algorithm used in our system for detecting the user's mood through facial expressions is the **Convolutional Neural Network (CNN)**. CNNs are particularly well suited for tasks involving image classification because of their capability to capture patterns present in images. The chosen architecture for this algorithm also consists of convolutional layers, pooling, flatten, and dense layers, which are designed to efficiently process and analyze the facial data to determine emotional states. This integration enables real-time mood assessment to tailor music recommendations more accurately. VGG16 model was also used for training to compare with the CNN model. The CNN model was also modified according to the optimizer function, number of epochs and learning rate to improve accuracy.

4.4 Model Creation

A CNN model is defined using the tf.keras.Sequential API. One of the convolutional layers have 32 filters with kernel size of (3,3) and ReLU activation. The pooling layer has a pool size of (2,2). The first dense layer uses ReLU activation and the second dense layer uses softmax activation for the multi-class Image classification. The optimizer used for the model is set to 'Adam'. The loss is calculated using crossEntropy. The ImageDataGenerator class from Tensorflow is used to perform Data augmentation and Normalization. The data is resized into (64,64) pixels. Then we train the model using the training and Validation data, for 10 epochs. This is the main model we have created to monitor the performance of the dataset. Further subsequent models were created with more epochs, different optimizers and different activation functions to know which suits best for the purpose so we use the model with the higher accuracy metric.

4.5 Model Performance

1. **Emotion Detection using Facial Recognition** The model that we created is implemented to perform real-time emotion detection of the user. The model that we created was saved and imported in as a pre-trained model here along with other dependencies like cv2 for video capturing and processing, numpy for numerical operations. The code initializes the video capture and uses the default camera of the devices considering the presence of one camera per device in use. A dictionary is defined, to map the indices of the predicted class to the corresponding emotion labels. After this the code enters into a loop to process the frames that get captured using the webcam feed. While it is inside the loop, it reads the frame by using cap.read function. The frame gets resized into (64,64) because even while we trained the model, we resized the image inputs into the same frame size. The predicted emotion label is laid on the frame and the position, thickness, font, font size, color are all specified. The frame is then displayed with the predicted emotion label. When the model was run with 10 epochs we got an increasing accuracy of 87%, so eventually on increasing the number of epochs to 50, the same lightweight model yielded us a frequency of 98.52%. Further models were created with changing the activation functions or changing the optimizer but the CNN model with 50 epochs gave us the best performance.

Table 1. Model Training Performance

Epoch	**Step Time (s)**	**Training Loss**	**Training Accuracy**	**Validation Loss**	**Validation Accuracy**
1	37s 41ms	1.6311	0.3631	1.5998	0.3820
5	36s 40ms	1.1192	0.5804	1.6635	0.4296
10	34s 38ms	0.6127	0.7847	2.3770	0.4082
15	30s 34ms	0.2585	0.9156	3.8773	0.4163
20	34s 38ms	0.1402	0.9572	5.3396	0.4152
25	35s 39ms	0.0883	0.9757	6.4961	0.4179
30	36s 40ms	0.0878	0.9751	7.1431	0.4140
35	36s 40ms	0.0746	0.9797	7.8096	0.4185
40	39s 43ms	0.0584	0.9842	7.9712	0.4087
45	35s 39ms	0.0546	0.9845	9.1647	0.4269
50	28s 31ms	0.0532	0.9852	9.0226	0.3954

2. **Music Recommendation system** Once the user's emotion is detected through the facial recognition system, this information is used to enhance our music recommendation system. The recognized emotional state of the user guides the selection of songs, creating a more personalized listening experience. This dynamic interaction ensures that the music recommendations align closely with the user's current mood, thereby enhancing user engagement and satisfaction. The combined dataset is trained and tested using various algorithms as seen in Table 2. The best algorithm SVM was picked and used for prediction. Using Spotify API the saved tracks of the user were obtained and the best algorithm was used to predict the output label mood. Spotify playlist was created for each of the moods using Spotify API.

Table 2. Model Evaluation Metrics

Model	Accuracy Score
Linear Discriminant	0.5616
K Neighbors	0.5228
Decision Tree	0.4650
Logistic Regression	0.5560
Support Vector Machine	0.5711
Gaussian Naive Bayes	0.5676

Table 3. Classification Report

Emotion	Precision	Recall	F1-Score	Support
Angry	0.57	0.60	0.58	90
Happy	0.60	0.75	0.67	153
Relaxed	0.53	0.66	0.59	125
Sad	0.62	0.28	0.39	129
Overall	**0.58**	**0.57**	**0.56**	**497**

4.6 Action based on Model output

Result When the user uses the application to predict their current emotion, the application has been designed in flask using the most accurate model we have created. The model was saved from the python notebook and then implemented in the basic structure of a flask app.

On running the flask app, the camera on the device takes 10 seconds to capture user emotion followed by the user clicking on the predicted emotion to listen to songs of that mood.

It is allowed for the user to click on the predicted emotion, complimentary emotion, or some other emotion entirely. The corresponding action of redirecting the

user to their selected emotion-based playlist is then performed to help the user enhance or improve their current mood.
The UI for this application was designed and kept basic and designed in HTML, CSS because of limited functionality of the features along with easy access to the user and better experience of using the web application.

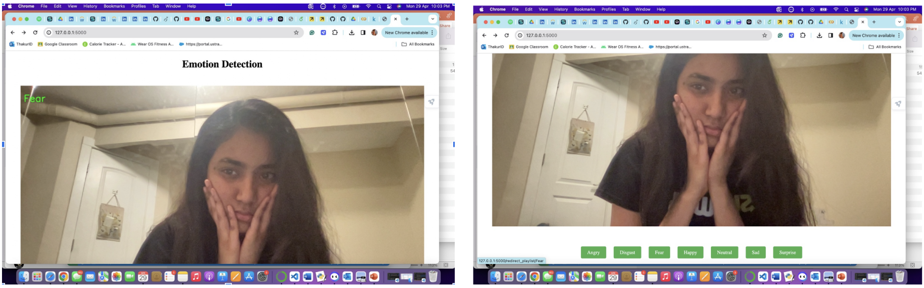


Fig. 3. The EmoTune flask application

5 Experiments

5.1 Experiment 1

The CNN model that we have created applies a convolutional operation with ReLU activation funcion, with a MaxPooling2D Layer and a Flatten layer and 2 Dense Layers with ReLU and Softmax activation respectfully. This model was then compared to a pre-trained model like VGG to compare and evaluate performance of model. The VGG model gave 99% accuracy while the CNN model gives us 98%.
The main point of observation is that even though the VGG16 model gives better accuracy on the dataset, the time taken to train the model is more than 6 days which is not feasible to use the model for further implementation because it will delay the application. This was also for just 10 epochs of the model.
The CNN model on the other side with 98% after running 50 epochs of the model in 20 minutes and gives almost the same accuracy while being a lightweight model too. This makes the model easy to save as the best performing model as compared to VGG16 model and then using it for the application. The observed trend is that the model even though it gives higher accuracy is not feasible to use and one of the reasons might be the complexity of the model.
For both these models, the system performance is evaluated using metrics like accuracy, precision, recall and F1-score. These were the evaluation metrics of choice because they will provide an overall assessment of the systems performance and consider both correct predictions as well as false positives and negatives. By comparing the systems performance with existing approaches, the aim

is to demonstrate superior accuracy and effectiveness in recognizing emotions from facial expressions.

5.2 Experiment 2

The analysis of the impact of different training data sizes on the performance of the emotion detection system will be tested using metrics like Accuracy and F1-Score. The main purpose of this experiment is to know about how performance of the system varies when it is trained on different subsets of the original dataset ranging from 10 to 100% of the total data. The selection of Accuracy and F1-score as evaluation metrics is because they are commonly used in classification tasks. This shows the robustness and the generalization capability of the model across different training data sizes which ultimately shows its practical applicability in scenarios with limited training data. We tried to scale down the data for training purposes from 28709 images to about 898 images to pass through the model. One of the reasons to do this is because since we had a very large dataset training to check which model performs better on the entire dataset was going to be difficult to keep up. Instead running the models with a scaled down dataset and then further deciding the model with the highest accuracy on the dataset as the final model will help us save time rather than loading around 30,000 images each time on a model. The shorter dataset however, did include all the categories of emotions to train from split equally.

5.3 Experiment 3

Another experiement we performed on the data was to change the activation function on the CNN model to Sigmoid and Softmax for the last dense layers and optimizer to rmsprop, the results on this however were not suitable to put to use at all. Considering that all the evaluation metrics we had stayed close to stagnant as the accuracy stayed at 25% for all 10 epochs of the model. One of the main reasons for this to happen might be the choice of activation function - Sigmoid for the model. sigmoid activation function is mainly used for binary classification tasks so it might not perform better on multi-class classification models like emotion recognition in this case.

While all the experiments have answered questions like which is the most suitable model, how much accuracy each model gives, there's still few unanswered questions about the models performance. About how accurately is it capturing information apart from the models performance but based on user input. Also to know about how the model functions, what is it considering to me primary feature - the lips, the eyes or eyebrows what is it. And in what time frame of capturing the user is emotion getting detected. Few more things can be taken care for for future work.

6 Conclusions

Overall, for the project we have implemented 2 working systems - SAong Recommendation based on mood and Emotion Recognition of the user. This multimodal application performs well with a final accuracy of 98% on the emotion detection model. This model is then implemented as a web app in flask, with a small UI in HTML, CSS made for user convenience.

While the application works well, there might be some ethical implications to the application about the frame or duration of emotion capture. Also what if the user wants a mood based playlist but doesnt want to show their face to the application. And is collecting user image as input an ethical way to solve the problem is a dilemma. However, the application does not store any user information as there is no database and everything is stored in the users own web session and will disappear when cache gets cleared. Overall, the application caters to a certain audience that would like to improve their listening experience when it comes to music based on their current mood.

References

1. Singh, J., Sajid, M., Yadav, C.S., Singh, S.S., Saini, M.: A novel deep neural-based music recommendation method considering user and song data. In: 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI). (2022) 1–7
2. Durga Malleswari, N.V., Gayatri, K., Sai Kumar, K.Y., Likhita, N., K, P., Bhattacharyya, D.: Music recommendation system using hybrid approach. In: 2023 Second International Conference on Electronics and Renewable Systems (ICEARS). (2023) 1560–1564
3. Chen, Y.: A music recommendation system based on collaborative filtering and svd. In: 2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS). (2022) 1510–1513
4. Sakti, S.M., Laksito, A.D., Sari, B.W., Prabowo, D.: Music recommendation system using content-based filtering method with euclidean distance algorithm. In: 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE). (2022) 385–390
5. Amara, K., Ramzan, N., Achour, N., Belhocine, M., Larbas, C., Zenati, N.: Emotion recognition via facial expressions. In: 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA). (2018) 1–6
6. Pandey, S., Handoo, S., Yogesh: Facial emotion recognition using deep learning. In: 2022 International Mobile and Embedded Technology Conference (MECON). (2022) 248–252
7. Chandraprabha, K.S., Shwetha, A.N., Kavitha, M., Sumathi, R.: Real time-employee emotion detection system (rteed) using machine learning. In: 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV). (2021) 759–763