Facial Image Based Emotion Detection and Music Recommendation System

Group No. 14
Akriti Agrawal (MT20021)
Saloni Gupta (MT20016)
Deepankar Kansal (MT20007)
Abhinay Gupta (2018209)
Sudeep Vig (MT20097)

ABSTRACT

Songs, as a medium of expression, have always been a popular choice to depict and understand human emotions. Reliable emotion based classification systems can go a long way in helping us parse their meaning. However, research in the field of emotion-based music classification has not yielded optimal results. In this paper, we are proposing a face based emotion recognition and music recommend system. The problem can be view as two different parts, first, detecting emotions from an input image which uses deep learning models to detect the emotion. Second, recommending music based on the detected emotion. In detecting emotion, our system has achieved accuracy of *65.58%* for seven emotion classes (angry, disgust, fear, happy, neutral, sad, surprise). While on four classes (happy, sad, angry, neutral) our model shows better accuracy of *73.58%*. In recommending music, the system uses mappings of emotions to music genres. These mappings helps in suggesting music to the user and feedback of the user can help as evaluation metric.

Keywords: Convolutional Neural Network (as CNN), Emotion recognition, Recommender system, OpenCV, VGG, ResNet50, InceptionV3.

1. INTRODUCTION

1.1 Motivation

Emotions play an essential role in identifying the mood of a human being. There are generally six raw emotions: happy, sad, anger, fear, surprise, disgust, contempt. Facial Expression Recognition with Music involves detecting human emotions and recommending the music that suits the user's mood. It is an automatic process that uses a person's face and diminishes hand interaction. Earlier, the system was only suggesting songs that were matching human emotion. In our approach, we will recommend soothing and calm songs if a person's mood is detected to be sad, angry, or fearful.

Current research work focus on the four main emotions named happy, sad, angry and neutral. We proposed the models that predict the seven classes named angry, disgust, fear, happy, neutral, sad and surprise. After recognizing the facial emotion for suggesting music, we have created playlist data to provide a link to famous YouTube playlists. For this task, mappings of emotions to the music genre have been made.

1.2 Problem Statement

In this project, we will propose a system which will take human's facial images as input and make predictions of emotions on that and after which it will suggest a list of songs which relates to the emotion detected of the input image as shown in Figure 1. We have seen that much good work had been done in the field of facial image's emotion detection, but in suggesting music there is no good approach so far, So we will try to apply different *Information Retrieval* techniques in suggesting or recommending the related music from a huge list of songs. In recognising images, in many research papers, they have applied Deep neural networks which acts as a black-box, So to better understand the facial expressions we will also try to apply a *Computer Vision* state-of-the-art to detect the emotions efficiently and easy to understand techniques. We will be using the two datasets for our project. The first dataset contains facial images available in Kaggle named as FER-2013, and another one is the songs dataset that we have created ourselves named as PLAYLIST DATA.

We see the results of the SVM (baseline), VGG-16, VGG-19 ResNet50 , InceptionV3 on the Kaggle dataset. Our baselines, SVM accuracy was 42.9% using PCA as feature extraction while VGG-16 had accuracy 63.3% , ResNet50 with accuracy of 42.61%, InceptionV3 with accuracy of 36.12%, VGG-19 with 57.63%. We observed most of these pre-defined models *overfits*, So after applying these standard models we tried lots of CNN variations to tackle overfitting problem and achieved an accuracy of *65.58%*.

Further section 2 briefly describes the past work done in the field. Section 3 presents our best CNN model architecture and other methodology. Finally, Sections 4 defines the baselines/models used for the emotion detection and music recommendation and the results obtained and evaluation metrics used for each.

2. LITERATURE REVIEW

In [1], they have used machine learning techniques and algorithms like Support Vector Machines (SVM), Neural Networks and Image preprocessing. In image preprocessing, they have done noise reduction, image to grayscale conversion, brightness enhancement, cropping and resizing. For

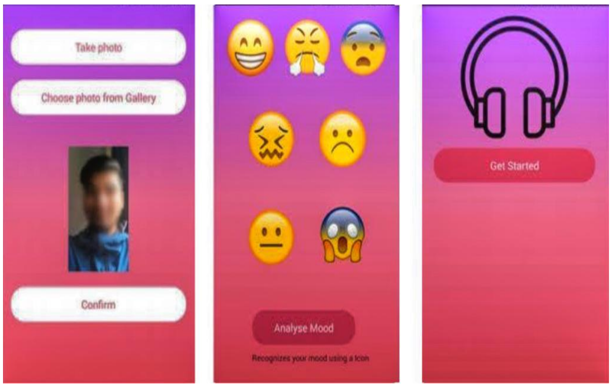


Figure 1: Problem Statement (Referred from this link)

feature extraction techniques like PCA (principal component analysis), LDA (linear discriminant analysis), ICA (independent component analysis) used. They applied machine learning and deep learning models like DBN (deep belief network) with RBM (Restricted Boltzmann Machine). For music recommendation, they have made folders for a particular emotion. To make the project more like real-life application, they have proposed a real-time Computer Vision using OpenCV, which takes the user's image. After preprocessing, emotion prediction is made using trained models and music is suggested. In image emotion recognition, they have achieved *72.4%* accuracy in which Happy emotion is detected with high accuracy of *92.6%*.

In [2], the authors have focused on mood recognition of OPM songs using lyrics.They have created the Word level features such as TF-IDF and keyGraph keyword generation algorithm, using different thresholds and parameters to determine how well these methods worked. They used two methods to label the mood of the songs: the manual annotation of songs and an automated approach using arousal and valence.After running the model, it was found that the manual approach performs better than the automatic approach.The keywords extracted from the lyrics that were manually labeled performs better. The keyGraph feature extraction gives an 80% average accuracy rating experimented on two different classification models were achieved. The study is limited as it gives good accuracy only for Sad Lyrics.For the automated annotation using arousal and valence values, more research is required for it to be a viable option for labeling songs.

A music player containing three main modules: Emotion Module, Music Classification Module, Recommendation Module, is proposed in [3]. They worked on four different classes happy, sad, angry and neutral, but the dataset used, contains seven emotion classes which leave a future aspect to work on all emotional classes. Emotion module achieved *90.23%* accuracy by applying CNN's on the images. OpenCV is also used for face detection before feeding to the network. The Music Classification module used the personalised dataset containing four classes that map to the original images data. Recommendation module does the mappings of classes of the other two modules. The user also has an option to select songs of his preferences from the detected emotion class folder. Overall they achieved *90.23%* accuracy in Emotion module, *97.69%* accuracy in correctly classifying the correct songs of emotions caught.

In [4], they have proposed an emotion-based music player. In this paper, the webcam is used to snap a person that will work as input for the system. A linear classifier is used to detect the face from the picture taken from the webcam. As soon as the face of the person is successfully detected, features extracted from the face is feed into PCA for dimension reduction, and particular emotion is detected with the use of multiclass SVM with a linear kernel is to compare the inputted data with stored one to see in what class (emotion) it belongs. After particular emotion is detected, the emotion is transferred to the web service. The song related to that emotion is played. Every song is assigned an emotion. Using webservice 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. The system is giving accuracy of around *90-95%*.

[5] proposed a work of facial emotion detection of 28 schizophrenia patients and 61 healthy people as subjects. Happy, sad, angry, fearful, disgusted, and neutral faces six are taken as classes of emotions. The intensity of emotions also taken into consideration. Observed that patients performed worse on fearful, disgusted and neutral expressions. They took eight low-intensity and eight high-intensity expressions of each emotion and 16 neutral expressions. When examined by intensity, recognition rates were lower in patients than comparison subjects for both low-intensity and high-intensity expressions. Out of all the emotions, happy faces were better recognized in different-gender photographs with an accuracy of 98%. Recognition was better for high-intensity than low-intensity expressions for all emotions except disgust, in which it is difficult to decide intensity.

[6] proposed a music information retrieval system from the user's reviews on the music and combining metadata and acoustic features related to that song. They have used Multimodal Album Reviews Dataset (MARD) dataset, which contains Amazon customer reviews and also applied mappings of MusicBrainz (MB) metadata and acoustic features from AcousticBrainz (AB) to the albums. After this, they have used bag-of-words (BoW) along with semantic(SEM) and sentiment(SENT) features. Training to test ratio made as 80-20% with 5-fold cross validation and using LinearSVM with BoW+SEM features noted the highest accuracy of *69.08%*.

[7] The focus is to recognize facial emotion and recommend songs efficiently. Music is the form of art that has a greater connection with a person's feelings. Artificial Intelligence is an eminent domain that has attracted many researchers. One of the star power of artificial intelligence is the face recognition technique. There are many existing systems recognizing facial emotions and many systems recommending music. Together these systems will recommend music based on facial emotion is the overall concept of paper. Few methodologies emphasized seven essential features that are common over age or different characters. They use OpenCV, especially the Adaboost algorithm, in the face recognition process. The face is recognized in complex colors of images

using a particular algorithm with the Adaboost algorithm. SVM is used as a primary characterization technique to order eight facial emotions. The faces distinguished utilizing channels in OpenCV changed to Greyscale. The image to be prepared acquired from a web camera or hard circle itself. Management and preparation have done using the "all versus one" approach of SVM to encourage multi-class characterization. The paper also utilizes Thayer's model of mindsets to perceive a music piece's state of mind. The feeling of a music piece is perceived via neural networks. Problems occur in recognizing facial emotion of unidentified elements like glasses, beard, etc.

In [8], they have proposed facial expressions recognition system using the facial action units. Facial action units are the facial muscle movements that describe a facial expression. They have performed this experiment on two datasets CK+ and OULU-CASIA Dataset.In this method, they have first obtained the facial action units then mapped them to the facial expression recognition using the classification methods. They have generated the facial action units using the two features, .i.e, appearance features and geometric features. Appearance features are obtained using the histogram's gradient, and geometric features have been obtained using the facial landmark points and facial expression alignment. After getting the facial action units, they have done max-min normalization to reduce irrelevant features. And then, at last, applied the three classification methods, i.e., Support vector regression(SVR), Extreme Gradient Boosting(Xgboost) and Deep neural nets(DNN). They have also used the 10-fold cross-validation in all the classifications. The result is that the CK+ gives better results than the OULU dataset because the CK+ dataset has good quality images. They have also used this method as a real-time system and have classified the images in about 30 ms. The conclusion is that if you want to use this method, it will only give the best results when you have the dataset with high-quality images.

3. METHODOLOGY

For our best model, we have given the input image of size $(48 * 48 * 1)$.

- In this model, firstly we have used the convolutional 2D layer with 32 filters with filter/kernel size of $(3, 3)$ and activation function ReLU.
- After that we have passed the output of this layer to the batch normalization layer.
- Then, we have again applied the convolutional 2D layer with 32 filters with filter/kernel size of $(3, 3)$ and activation function ReLU.
- After that, we again applied the batch normalization.
- Then, we applied the max-pooling layer with filter size$(2, 2)$.
- Then, we applied dropout layer.

After that, we applied the same layers, but we changed the number of filters from 32 to 64 and then from 64 to 128. Similarly, we applied pooling and dropout layers.

At the end of the model, we first used the flatten layer. After this, we have used the dense layer/ fully connected layer of size 128, with activation function ReLU. Then, we applied the batch normalization. Then, we used the dropout layer. At the last, we have used the dense layer of size 7, with activation function softmax.
To better understand this model, please have a look at following figure 2.

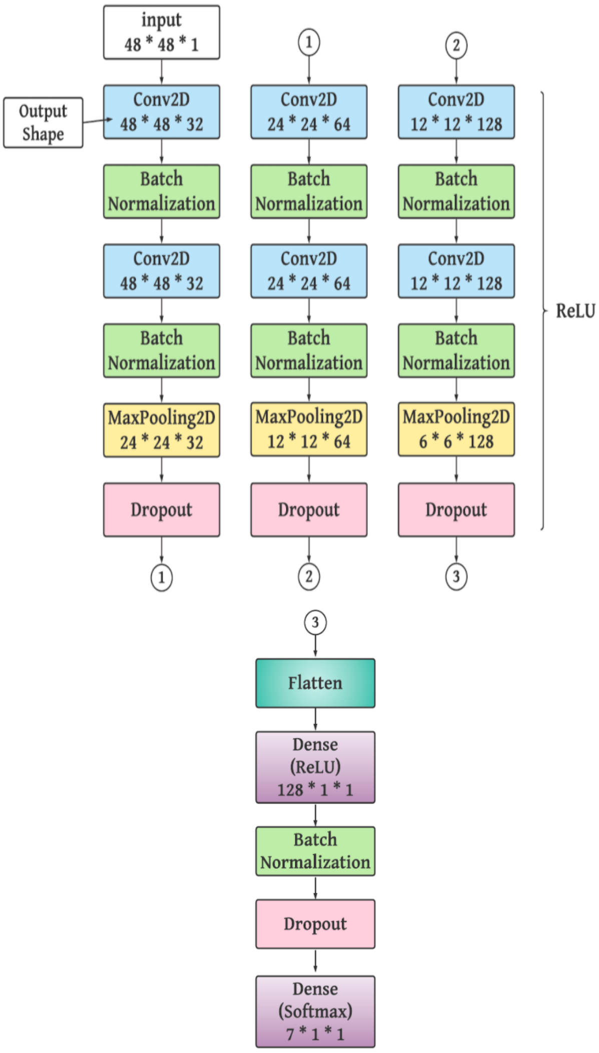


Figure 2: CNN architecture used for training on facial images

In the second model, we have given the input image of 48*48*1.
This model is almost similar to the best model but having a different number of filters in the Convolutional 2D layer and also having the difference while using the dropout layer. We have used the:

- Filters of sizes 16,32 and 64 at the first, second and third layer, respectively.
- Dropout of 0.25 at the first use and 0.5 at the rest of

the layers.

The best model has a dropout of 0.3 at the first layer and 0.5 at the rest of the layers. This model gives an accuracy of 64% and also has reduced overfitting.

Similarly, the third model is having the :

- Filters of sizes 128,256, and 512 at the first, second and third layer, respectively.
- Dropout of 0.25 at all the layers.

This model is having an accuracy of 61%.

In the fourth model, we have added one more layer. In this model, we have:

- Filters of sizes 16,32,64 and 128 at the first, second, third and fourth layer, respectively.
- Dropout of 0.25, 0.3 at first and second layers and 0.5 at rest of the layers.

This model is having an accuracy of 64% but having more overfitting than the second model.

In the fifth model, we have one more layer than the fourth model. In this, we have:

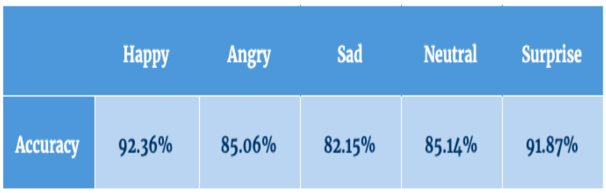
- Filters of sizes 16,32,64,128 and 256 at the first, second, third, fourth and fifth layer, respectively.
- Dropout of 0.25, 0.3 at first and second layers and 0.5 at rest of the layers.

This model gives an accuracy of 63% and having less overfitting than the fourth model but more than the second model.

After trying various CNN variants, we follow one of the approaches discussed in previous research papers which is, to classify the data on four emotion classes; *angry, happy, sad, neutral.* We used our best CNN variant and got accuracy of *73.58%*, which is outperforming the previously reported accuracy of *72.4%* discussed by Ananya Dhar and Bilal N. Shaikh [1]. We also tried *under-sampling* of the data, but proportions of these four emotion classes are good, So, under-sampling the data produces very similar results.

Ensemble Method
After getting *65.58%* accuracy for 7 classes and *73.58%* accuracy for 4 classes, We tried ensemble method for 5 emotions; angry, happy, sad, surprise, neutral. In this approach, first dataset is created for all the emotion classes separately, i.e. 5 datasets to contain only two labels 0 and 1, 1 for the emotion class for which we are creating the model and 0 for other classes. Using this approach, we trained 5 models for specified emotion classes. Accuracies reported by different emotion class model are presented in figure 3.

When we calculate overall accuracy after combining these models by predicted probabilities criteria, accuracy comes out to be ***75.3%***, which is outperforming the baselines as well as CNN variants, as this is calculated for 5 emotion classes whereas CNN is giving 73.58% for 4 emotion classes.



	Happy	Angry	Sad	Neutral	Surprise
Accuracy	92.36%	85.06%	82.15%	85.14%	91.87%

Figure 3: Accuracy Of Different Emotion Classes

4. RESULTS & EVALUATION

4.1 Evaluation Metrics

4.1.1 Accuracy

It is the number of correct predictions made by the model over all the predictions. Accuracy is a good measure when the target variable class is balanced. Accuracy is not preferred as evaluation metric when one target variable class is in majority.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Total number of samples}}$$

4.1.2 F1-score

It is the harmonic mean of Precision and Recall. It gives a better measure of incorrectly classified cases than accuracy. It is a better evaluation metric in case of imbalanced classes.

$$\text{F1-score} = 2 * \frac{\text{Precision} + \text{Recall}}{\text{Precision} * \text{Recall}}$$

where Precision and Recall are defined as below.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+ False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+ False Negative}}$$

4.1.3 Confusion Matrix

It is used for classification problems for finding the correctness and accuracy of the model. Terms associated with Confusion matrix:
1. True Positive: When both the actual class and the predicted class is 1(True).
2. True Negative: When both the actual class and the predicted class is 0(False).
3. False Positive: When the actual class is False but the predicted class is True.
4. False Negative: When the actual class is True but the predicted class is False.

4.2 Baseline Results

We achieved an overall classification accuracy of 65.58% (on the testing data) on seven emotion classes using Convolutional Neural Network (CNN), as described in figure 2. Although we also tried various other pre-defined models like VGG-16, VGG-19(accuracy of 57.63% on the testing data and accuracy of 97.53% on training data), InceptionV3(accuracy of 33.54% on the testing data and accuracy of 36.12% on training data), ResNet50(accuracy of 42.61% on the testing data and accuracy of 94.33% on training data), XceptionNet(accuracy of 38.90% on testing data) and a machine learning model SVM (Support Vector Machine). But neither of them is giving good accuracy, or some models like VGG-16, VGG-19 and ResNet50 overfits the data alot, emaple of VGG-19 is described in figure 4.

For music recommendation, various research papers are using their own datasets and calculating accuracy according to this data. So, comparing recommender system's results are out of the scope of this project and we also provide our own recommender system.

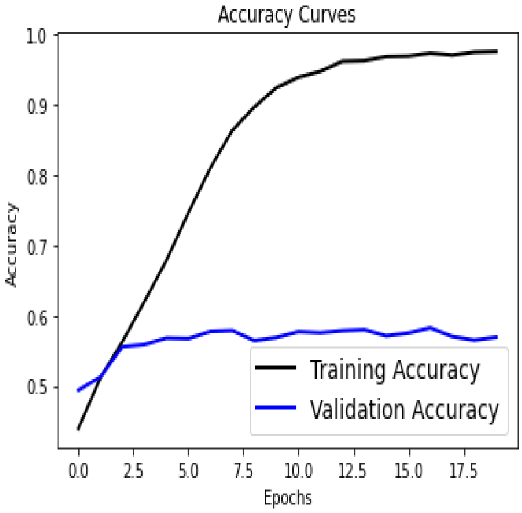
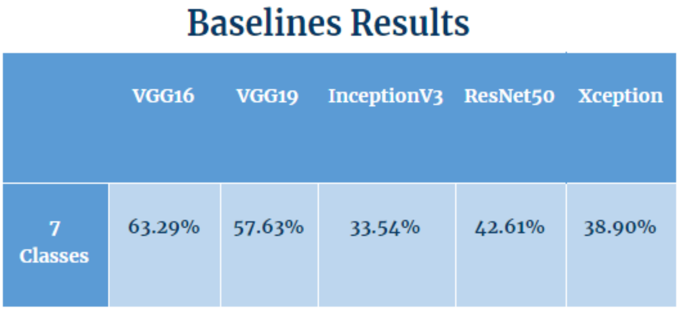


Figure 4: Accuracy Vs Epochs plot for VGG-19 model


Baselines Results

	VGG16	VGG19	InceptionV3	ResNet50	Xception
7 Classes	63.29%	57.63%	33.54%	42.61%	38.90%

Figure 5: Accuracy Of Different Baselines

4.3 Proposed Method Results

4.3.1 Emotion Recognition

So, we tried variations of CNN (as described in papers) and came to the one described in figure 2. The proposed CNN architecture gives 65.58% accuracy and 0.656 F1-score on testing data for seven emotion classes and 74.03% accuracy on the training data. We can say that this model also overfits a bit! Therefore, we are also submitting one more CNN variation, which gives 64.07% accuracy and 0.641 F1-score on the testing data and 67.26% accuracy on the training data for seven classes. This accuracy is slightly lesser than the proposed model, but the problem of overfitting addressed well. The accuracy plot for this model is presented in the figure 6.

After seven class classification, we tried our best CNN variant for four class classification and got 73.58% accuracy which is outperforming the previously reported accuracies for four class classification. We also tried *ensemble learning on five classes*, and achieved 75.3% accuracy when calculating on happy, angry, sad, surprise and neutral emotion classes.

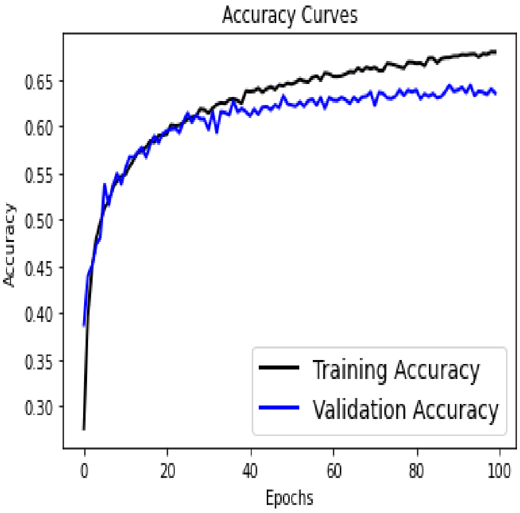
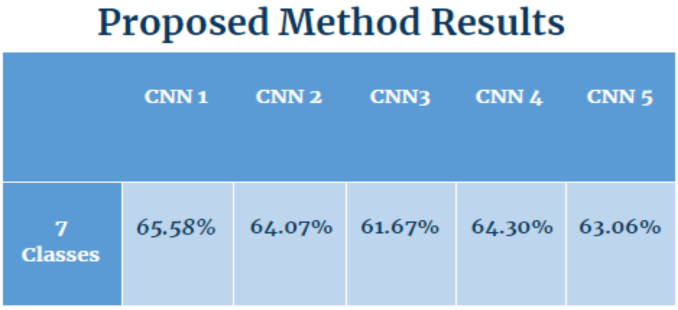


Figure 6: Accuracy Vs Epochs plot for CNN model


Proposed Method Results

	CNN 1	CNN 2	CNN3	CNN 4	CNN 5
7 Classes	65.58%	64.07%	61.67%	64.30%	63.06%

Figure 7: Accuracy Of Different CNN Variants

4.3.2 Recommender System

After detecting emotion, we suggest a playlist of songs according to emotion detected, for which we are using PLAYLIST DATA containing 100 top rated playlists from YouTube and other platforms according to genres. For suggesting music, we map some genres to each emotion. These mappings were made according to emotion, i.e. for sad emotion, calm and soothing song genres are mapped like folk, rock, pop. Further, *intensity* predicted by the emotion recognition module is also taken into account and mapped genres to emotion according to intensity predicted. One can take feedback from the user for betterment of the system.

5. CONCLUSION

Results obtained are very promising. The system has high accuracy and quick response time which makes it suitable for most practical purposes. Using *OpenCV* module to extract faces and then detecting emotions, makes the system more robust for specific data. Proposed CNN architectures giving accuracy of *65.58%* in the case of seven class emotion recognition and *73.58%* accuracy in the case of four class emotion recognition. While recommending music, mappings of emotion to music genre helps a lot. These mappings can be modified based on the feedback received from the user and thus increase the system's performance to a well extent.

We observed that model predicts very few emotions correctly when tried on unseen data if the data is not well pre-processed. To achieve this goal, *ensemble technique* is helpful where one can train one emotion class against all other classes and finally predict based on all produced models. We tried this approach, and found that model's accuracy increased to a great extent and giving *75.3% accuracy and 0.753 F1-score* for five class classifications.

Future aspects for this system will be, to enhance accuracy on unseen data by training model for specific emotion classes and then tune them in a way, so that model works fine for unseen data. Also one can improve, music recommender system by taking feedback from the user in a more constructive way.

6. REFERENCES

[1] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[2] Ananya Dhar and Bilal N. Shaikh Mohammad. Emotion Recognition with Music using Facial Feature Extraction and Deep Learning. SSRN Scholarly Paper ID 3560840, Social Science Research Network, Rochester, NY, April 2020.

[3] Rafael Cabredo Emil Ian V. Ascalon. Emotion based music recommendation system. *Lyric-Based Music Mood Recognitio*, 3:1–8, 2015.

[4] Shlok Gilda, Husain Zafar, Chintan Soni, and Kshitija Waghurdekar. Smart music player integrating facial emotion recognition and music mood recommendation. In *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 154–158. IEEE, 2017.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] H Immanuel James, J James Anto Arnold, J Maria Masilla Ruban, M Tamilarasan, and R Saranya. Emotion based music recommendation system. *EMOTION*, 6(03), 2019.

[7] Christian G Kohler, Travis H Turner, Warren B Bilker, Colleen M Brensinger, Steven J Siegel, Stephen J Kanes, Raquel E Gur, and Ruben C Gur. Facial emotion recognition in schizophrenia: intensity effects and error pattern. *American Journal of Psychiatry*, 160(10):1768–1774, 2003.

[8] Sergio Oramas, Luis Espinosa-Anke, Aonghus Lawlor, Xavier Serra, and Horacio Saggion. EXPLORING CUSTOMER REVIEWS FOR MUSIC GENRE CLASSIFICATION AND EVOLUTIONARY STUDIES. *New York City*, page 7, 2016.

[9] Deny John Samuvel, B. Perumal, and Muthukumaran Elangovan. Music recommendation system based on facial emotion recognition. *3C Tecnología. Glosas de innovación aplicadas a la pyme*, pages 261–271, March 2020.

[10] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[11] Jiannan Yang, Fan Zhang, Bike Chen, and Samee U. Khan. Facial Expression Recognition Based on Facial Action Unit. In *2019 Tenth International Green and Sustainable Computing Conference (IGSC)*, pages 1–6, Alexandria, VA, USA, October 2019. IEEE.