Models for Emotion Detection and Music Recommendation System using SR-GAN

DR. LATHA H N[1], DR. GAYATHRI M S[2], DR. KIRAN BAILEY[3]
[1, 3] Dept. Electronics and Communication, BMS College of Engineering, Basavanagudi, Bangalore
[2] Dept of Mathematics, BMS College of Engineering, Basavanagudi, Bangalore

Abstract— Emotions are an important part of our life. A living being cannot live without emotions. The emotion of a person is affected by many things in and around it. A person's likes and dislikes vary according to his/her current emotion. Hence if we know the mood of the person we could recommend him/her the products that he would like to have during that situation. In this paper we have proposed a system which predicts the facial emotion of the person and then recommends a song to the person. We have used the FER2013 [9] dataset for this purpose and have used different techniques for data generation, Image Super Resolution, and classification. A Fine Tuned Swin transformer model was used for classification. There were seven classes (emotions) in the dataset. The model could achieve an accuracy of 66.8% on the test-set with a best recall of 0.92 for disgust and happy classes and lowest recall of 0.29 for fear class.

Index Terms- Classification, Recommendation, Emotion, Generative Adversarial Network, Transformers

I. INTRODUCTION

Emotions [1] are mental states which depend on neurophysiological changes and it is variously associated with thoughts, feelings, and behavioral responses. There are various Human emotions, sad, disgust, anger, fear, surprise, neutral and happy are few of its classifications. Emotion detection plays a very important role in our life since people express their emotions in different ways and this leads to decisions and choices taken. Many factors have an influence on the emotion like the people around us virtually [2] or physically, music we listen to, things that happen to us or in and around us. There are different types of emotions [3] like happy, sad, surprise, angry, etc. Emotions also define how we behave [4].

Music industry is growing day by day. The main reason behind this is that music is assumed to be a great healer and it tends to improve one's mood. It can be also used as a soul soother. The music industry is continuously trying to make the people connect with the music which they produced and in continuation with that it also tries to analyze the different emotions of various people. Music on the other hand plays an important role in each and every function so it can be assumed to be a bare minimum of every function. So if an algorithm is developed such that it suggests music according to the person's mood it can be a boon in the music sector.

Likewise, paying attention to the perfect sort of music at the ideal time might work on mental health [5]. In this manner, human feelings have a solid relationship with music [6].

Our work focuses on merging emotion detection and music in order to improve the quality of life of people. The work involves the usage of deep learning concepts and the creation of a real time application, either using a laptop and webcam, or as a mobile application.

II. RELEVANT WORK

In the past work has been done in the emotion detection sector where a machine learning model is used to detect the emotion of a particular person. In this paper [7] the author discussed the various features of emotion and how they are related to the different moods of humans. They have classified various emotions and then categorized them into different heads like disgust etc. They also detect human activities which are controlled by the neural system of the body i.e., brain. In this way all human activities can be mapped to different moods.

Similarly machine learning is used to recommend songs to the respective person based on either text or emotion. In this paper [8] the author discusses various techniques to recommend a song to the user. With the rapid expansion of format used to store the music, searching and storing music has become very important. However music data recovery methods have been made effectively over the most recent decade, the improvement of music recommender frameworks is currently at a beginning phase.. There are two famous calculations which are utilized for music proposal systems, that are cooperative separating and content-based models, have been found to perform well, But there is a problem in those methods that is the relatively poor experience due to the inability to find the music which contains emotional meaning in the music. So the author has given two of user centered approaches, they are a context-based model and emotion-based model.

III. DATA-SET

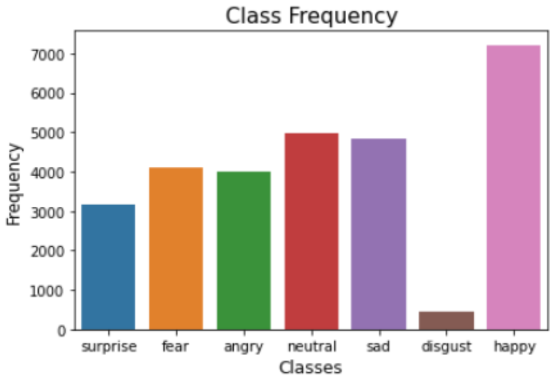


Figure 1: Dataset Distribution

The dataset used in the work was FER2013 [9]. FER full form is Facial Emotion Recognition. This dataset was released in the year 2013 and is widely used for building facial emotion recognition models. The dataset consists of 28709 face images with seven kinds of emotions namely happy, sad, disgust, angry, neutral, fear, and surprise. This is an imbalanced dataset with class “happy” having the highest number of images with count 7215 and class “disgust” which has the lowest number of images with count that is 436 images. This dataset was then divided into 60% train, 20% validation, and 20% as test-set.

IV. METHODOLOGY

In this section we would be explaining in brief the models and techniques we have used and the system that predicts the mood of the person based on the emotion of the input face image.

a. Data Generation:

We have used the dataset from the face emotion recognition 2013 for our work. We had 3900+ images for angry, sad, happy, fear, neutral images but there were only 436 images for disgust images. To generate [10] more disgust images we have used a generative adversarial network. Generative adversarial network [11] is a machine learning framework which uses deep learning networks and is used to generate images with similar features to the input data. It contains a discriminator network and a generator network. The generator generates fake images using the input real images and tries to fool the discriminator; Discriminator tries to distinguish between the real images and fake images. In this process both networks get trained and become efficient in what they are doing. This will help us to balance the images; this will also help us to get good results.

b. Image Super Resolution:

Image super resolution refers to the task of increasing the resolution of the image from lower resolution to higher resolution. Thereby we will increase the quality of the image without the loss of the features of the image. Models like MobileNet [12] require images of size 244X244X3. If we try to maximize the image to specific resolution using convenient techniques, we will lose a lot of quality and features of the image. Image super resolution uses GAN to enhance the resolution of the image[13]. The particular model we are using will help us to increase the resolution by four times. Increasing the resolution using Image super resolution will help us to retain the important features of the image which will help us to improve the model. Figure 2 depicts the results of the image super resolution.

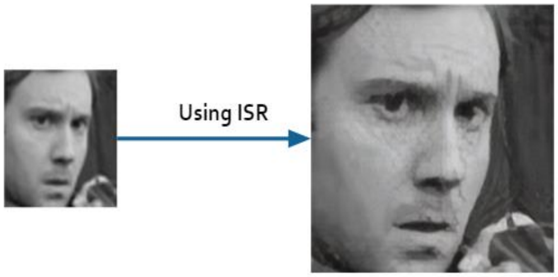


Figure 2: Image Super Resolution

c. Classification

Classification [14] refers to labeling the input to a particular class based on the characteristics of the input. Classification models are used for this task. Classification is the most important and widely researched topic which led to the creation of models that are highly accurate and fast. In our system we have used the Swin Transformer (SwinT) model to classify the images. The inputs to the classification model are the super resolution images. The SwinT is a Vision Transformer [15]. SwinT [16] is transformer based backbone architecture for visual tasks and the first one to be so. Compared to the words in the text the images have large variations in the pixel count and also the scales of objects in the image are the problems which were addressed by SwinT which led to getting the best results.

d. *Proposed System*

The system we have proposed takes in an image and plays music depending on the detected emotion of the face in the image. As a first step we have generated the disgust images to reduce the effect of the class imbalance. The images of 48x48 were processed to increase their size to 224x224x3 using super resolution technique. These images were then used to train the classification model. Once the model is trained this model can be used in production.

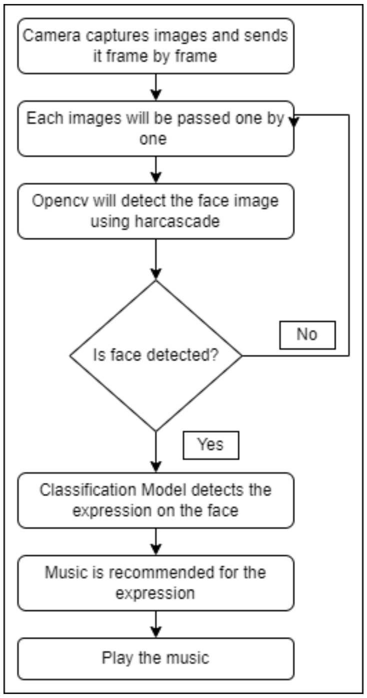

Camera captures images and sends it frame by frame
Each images will be passed one by one
Opencv will detect the face image using harcascade
Is face detected?
No
Yes
Classification Model detects the expression on the face
Music is recommended for the expression
Play the music

Figure 3 : Process Flowchart

In production the frames with some frequency from the camera are sent to the Haar cascade to detect the face of the person, if the face is not detected the frame is discarded, else if it is detected this face image is then passed to the super resolution model. The classification model then detects the emotion, this information is sent to the music recommendation subsystem. This subsystem based on the face emotion randomly selects a song to play from the pool of songs for that particular emotion. The music player plays the music. The flowchart of the process is depicted in figure 3.

V. RESULTS

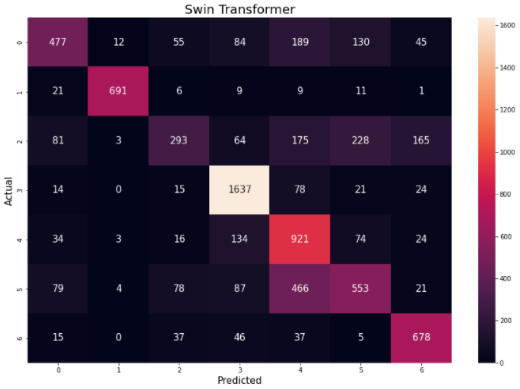


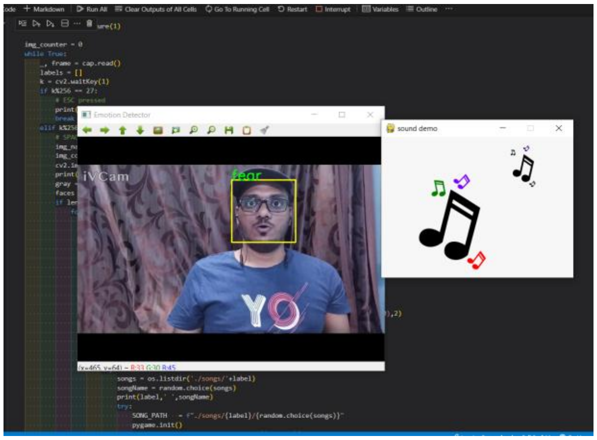
Figure 4: Confusion Matrix

This section throws a light on the results that we have obtained during the training and testing phase of the classification model. We have used the Pytorch library to implement the model. A transformer based model called SwinT was used for the classification. The input image was classified into one of the 7 classes. The classes were namely angry, disgust, fear, happy, neutral, sad, and surprise with labels from 0 to 6 respectively. Accuracy was used as the metric [17] to track the model performance and validation loss was used as the decider if training had to be continued or stopped. This technique is called early stopping [18] that prevents the model from overfit. Overfit is a situation in which the model performs well on the train set but poorly on the dataset that it has not seen during the training phase [19]. The training of the model was done on 11,000 images and tested on 7850 images. Our classification model could achieve validation accuracy of 68% and test accuracy of 66.8%.

Table 1: Validation accuracy and loss of CNN and SWINT models

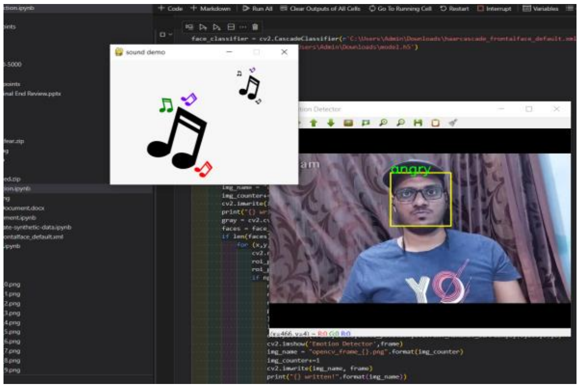
Model	Validation accuracy (%)	Validation loss
CNN model	55	1.89
SWINT model	68%	1.07

The performance of the Swin transformer model on each class of test set is shown in the figure 4. We can see that the model has performed the best for class “disgust” and “happy” with a recall of 0.92. Infact the available FER2013 dataset had very less “disgust” images. Hence generating new images of that class have improved the predictions. Also the model has performed poorly on classes “fear” with recall of 0.29. The model has misunderstood “sad” images as “neutral”. This is evident as 466 sad images were predicted as neutral. This may be due to some “sad” images having very slight changes in the expression from the neutral position. Hence this made the class “neutral” to have the lowest precision of 0.49. The weighted average F1-score, precision, and recall are 0.65, 0.67, and 0.67 respectively.



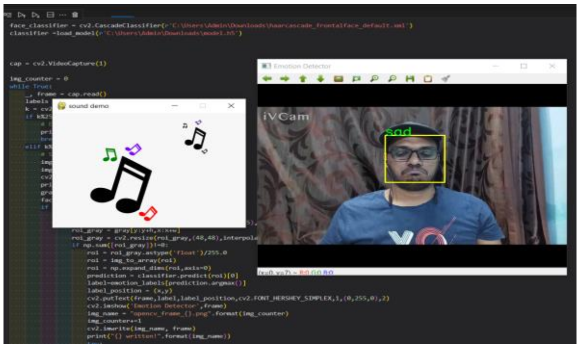
```
ure(1)

img_counter = 0
while True:
    _, frame = cap.read()
    labels = []
    k = cv2.waitKey(1)
    if k%256 == 27:
        # ESC pressed
        print(
        break
    elif k%256
        # SPA
        img_n
        img_c
        cv2.im
        print(
        gray
        faces
        if le
            fo
                                                        ),2)
                songs = os.listdir('./songs/'+label)
                songName = random.choice(songs)
                print(label,' ',songName)
                try:
                    SONG_PATH   = f"./songs/{label}/{random.choice(songs)}"
                    pygame.init()
```

Figure 5: Detection and recommendation for fear emotion



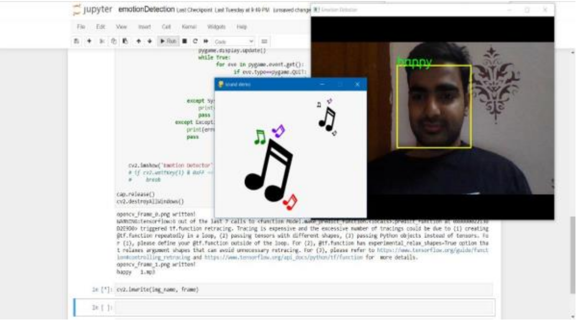
```
face_classifier = cv2.CascadeClassifier(r'C:\Users\Admin\Downloads\haarcascade_frontalface_default.xml
                                   sers\Admin\Downloads\model.h5')

        img_name =
        img_counter+
        cv2.imwrite(
        print("{} wr
        gray = cv2.c
        faces = face
        if len(faces
            for (x,y
                cv2.r
                roi_
                roi_
                if np
                    cv2.imshow('Emotion Detector',frame)
                    img_name = "opencv_frame_{}.png".format(img_counter)
                    img_counter+=1
                    cv2.imwrite(img_name, frame)
                    print("{} written!".format(img_name))
```

Figure 6: Detection and recommendation for Angry emotion



```
face_classifier = cv2.CascadeClassifier(r'C:\Users\Admin\Downloads\haarcascade_frontalface_default.xml')
classifier =load_model(r'C:\Users\Admin\Downloads\model.h5')


cap = cv2.VideoCapture(1)

img_counter = 0
while True:
    _, frame = cap.read()
    labels
    k = cv2
    if k%25
        # E
        pri
        bre
    elif k%
        # S
        img
        img
        cv2
        pri
        gra
        fac
        if
                                                   5),
            roi_gray = gray[y:y+h,x:x+w]
            roi_gray = cv2.resize(roi_gray,(48,48),interpola
            if np.sum([roi_gray])!=0:
                roi = roi_gray.astype('float')/255.0
                roi = img_to_array(roi)
                roi = np.expand_dims(roi,axis=0)
                prediction = classifier.predict(roi)[0]
                label=emotion_labels[prediction.argmax()]
                label_position = (x,y)
                cv2.putText(frame,label,label_position,cv2.FONT_HERSHEY_SIMPLEX,1,(0,255,0),2)
                cv2.imshow('Emotion Detector',frame)
                img_name = "opencv_frame_{}.png".format(img_counter)
                img_counter+=1
                cv2.imwrite(img_name, frame)
                print("{} written!".format(img_name))
```

Figure 7: Detection and recommendation for Sad emotion

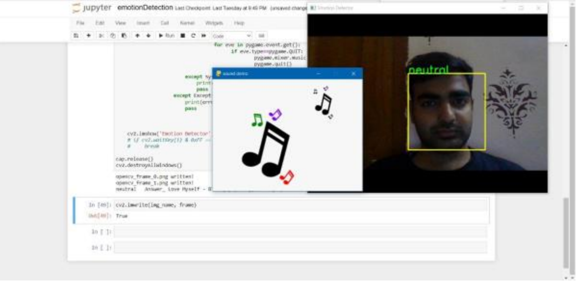

jupyter emotionDetection Last Checkpoint: Last Tuesday at 9:49 PM (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help
Run
Code
Emotion Detector
sound demo
happy

```
                    pygame.display.update()
                    while True:
                        for eve in pygame.event.get():
                            if eve.type==pygame.QUIT:

                except Sy
                    print(
                    pass
            except Except
                print(err
                pass

    cv2.imshow('Emotion Detector'
    # if cv2.waitKey(1) & 0xFF ==
    #     break

cap.release()
cv2.destroyAllWindows()
```

```
opencv_frame_0.png written!
WARNING:tensorflow:6 out of the last 7 calls to <function Model.make_predict_function.<locals>.predict_function at 0x
D2E9D0> triggered tf.function retracing. Tracing is expensive and the excessive number of tracings could be due to (1) creating
@tf.function repeatedly in a loop, (2) passing tensors with different shapes, (3) passing Python objects instead of tensors. Fo
r (1), please define your @tf.function outside of the loop. For (2), @tf.function has experimental_relax_shapes=True option tha
t relaxes argument shapes that can avoid unnecessary retracing. For (3), please refer to https://www.tensorflow.org/guide/funct
ion#controlling_retracing and https://www.tensorflow.org/api_docs/python/tf/function for  more details.
opencv_frame_1.png written!
happy   1.mp3
```

```
In [*]: cv2.imwrite(img_name, frame)

In [ ]:
```

Figure 8: Detection and recommendation for Happy emotion


jupyter emotionDetection Last Checkpoint: Last Tuesday at 9:45 PM (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help
Run
Code
Emotion Detector
sound demo
neutral

```
                        for eve in pygame.event.get():
                            if eve.type==pygame.QUIT:
                                pygame.mixer.music
                                pygame.quit()

                except Sy
                    print(
                    pass
            except Except
                print(err
                pass

    cv2.imshow('Emotion Detector'
    # if cv2.waitKey(1) & 0xFF ==
    #     break

cap.release()
cv2.destroyAllWindows()
```

```
opencv_frame_0.png written!
opencv_frame_1.png written!
neutral   Answer_ Love Myself -
```

```
In [49]: cv2.imwrite(img_name, frame)
Out[49]: True

In [ ]:

In [ ]:
```

Figure 9: Detection, recommendation for Neutral emotion

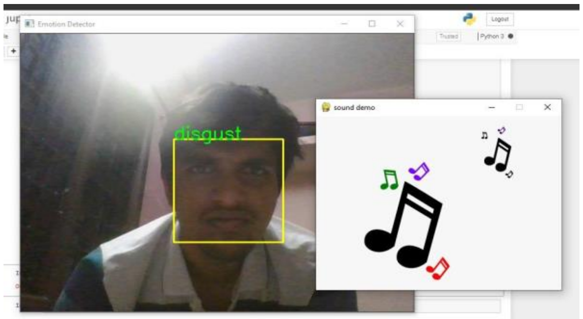


Figure 9: Detection and recommendation for Disgust emotion

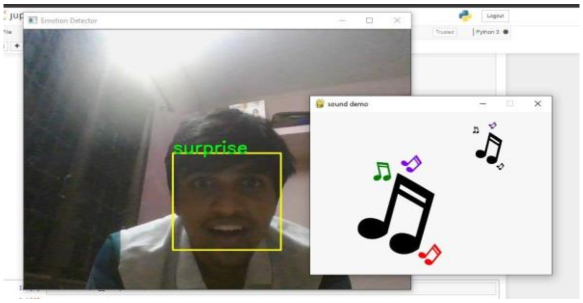


Figure 10: Detection and recommendation for Surprise emotion

CONCLUSION

The FER2013 dataset used had images consisting of white faces only. The data generated using GANs showed positive results in the performance of the model in detecting "disgust" class. Image super resolution increased the accuracy of the model when compared to normal resizing. The model could not perform well in all the classes. Performed the best on class "disgust" and "happy" and performed poorly on class "fear". As said already, since the dataset had only white faces the model will not perform well when the input image has a brown or a black face. Hence the dataset could be diversified by including the faces from different parts of the world as people at different regions will have different characteristics for the face. We could also try different models and techniques for classification of the image with finetuning.

REFERENCES

[1] Izard, Carroll E. "Emotion theory and research: highlights, unanswered questions, and emerging issues." Annual review of psychology vol. 60 (2009): 1-25. doi:10.1146/annurev.psych.60.110707.163539

[2] Utz, Sonja. (2019). Social Media as Sources of Emotions. 10.1007/978-3-030-13788-5_14.

[3] Cowen, Alan. (2018). How Many Different Kinds of Emotion are There?. Frontiers for Young Minds. 6. 10.3389/frym.2018.00015.

[4] Edward A. Selby, Michael D. Anestis, Thomas E. Joiner, Understanding the relationship between emotional and behavioral dysregulation: Emotional cascades, Behaviour Research and Therapy, Volume 46, Issue 5, 2008, Pages 593-611, ISSN 0005-7967, https://doi.org/10.1016/j.brat.2008.02.002.

[5] Hallam, Susan. (2010). The power of music: Its impact on the intellectual, social and personal development of children and young people. International Journal of Music Education. 28. 269-289. 10.1177/0255761410370658.

[6] Swaminathan, Swathi & Schellenberg, E.. (2015). Current Emotion Research in Music Psychology. Emotion Review. 7. 189-197. 10.1177/1754073914558282.

[7] Singh, Dilbag. (2012). Human Emotion Recognition System. International Journal of Image, Graphics and Signal Processing. 4. 10.5815/ijigsp.2012.08.07.

[8] Song, Yading & Dixon, Simon & Pearce, Marcus. (2012). A Survey of Music Recommendation Systems and Future Perspectives.

[9] L. Zahara, P. Musa, E. Prasetyo Wibowo, I. Karim and S. Bahri Musa, "The Facial Emotion Recognition (FER-2013) Dataset for Prediction System of Micro-Expressions Face Using the Convolutional Neural Network (CNN) Algorithm based Raspberry Pi," *2020 Fifth International Conference on Informatics and Computing (ICIC)*, 2020, pp. 1-9, doi: 10.1109/ICIC50835.2020.9288560.

[10] Singh, Aditya & Dutt, Varun. (2020). Data Generation From Random Noise with Generative Adversarial Networks (GANs).

[11] Goodfellow, Ian & Pouget-Abadie, Jean & Mirza, Mehdi & Xu, Bing & Warde-Farley, David & Ozair, Sherjil & Courville, Aaron & Bengio, Y.. (2014). Generative Adversarial Networks. Advances in Neural Information Processing Systems. 3. 10.1145/3422622.

[12] Howard, Andrew & Zhu, Menglong & Chen, Bo & Kalenichenko, Dmitry & Wang, Weijun & Weyand, Tobias & Andreetto, Marco & Adam, Hartwig. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.

[13] Hemanth, K., and H. N. Latha. "Dynamic scene Image deblurring using modified scale-recurrent network." 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE, 2020.Song,

[14] Shachee, S. B., H. N. Latha, and N. Hegde Veena. "Electrical energy consumption prediction using LSTM-RNN." Evolutionary Computing and Mobile Sustainable Networks: Proceedings of ICECMSN 2021. Singapore: Springer Singapore, 2022. 365-384.

[15] Latha, H. N., and Rajiv R. Sahay. "A local modified U-net architecture for image denoising." *reconstruction* 8 (2020): 14.

[16] Latha, H. N., and S. Ramachandran. "Design Of Context Based Adaptive Variable Length Coding And Deblocking Filter for H. 264." Procedia Technology 4 (2012): 671-676.

[17] Latha, H. N., and Bharathi Lokesh. "Denoising and deblurring by gauss markov random field: an alternating minimization convex prior." International Journal of Science and Research (2019): 1669-1672.

[18] Shylaja, H. N., H. N. Latha, and H. N. Poornima. B Uma "Detection and Localization of Mask Occluded Faces by transfer learning using Faster RCNN." Available at SSRN 3835214 Elseviour, March 2021.

[19] Vilohit Tapashetti, Vaishnavi S, N Latha H, Vijaya K, 2023/6/26 Quest Journals Journal of Software Engineering and Simulation Volume 9 Issue 6 (2023), pp: 24-30, Publisher, Quest Journals

[20] Yading & Dixon, Simon & Pearce, Marcus. (2012). A Survey of Music Recommendation Systems and Future Perspectives.

[21] Tammina, Srikanth. (2019). Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images. International Journal of Scientific and Research Publications (IJSRP). 9. p9420. 10.29322/IJSRP.9.10.2019.p9420.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, et.al “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.“ ICLR 2021

[23] Liu, Ze., Lin, Yutong., Cao, Yue., Hu Han., Wei, Yixuan., Zhang, Zheng., Lin, Stephen., Guo, Baining. 2021. “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows”

[24] Hossin, Mohammad & M.N, Sulaiman. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process. 5. 01-11. 10.5121/ijdkp.2015.5201.

[25] Prechelt, Lutz. (2000). Early Stopping - But When?. 10.1007/3-540-49430-8_3.

[26] Ying, Xue. (2019). An Overview of Overfitting and its Solutions. Journal of Physics: Conference Series. 1168. 022022. 10.1088/1742-6596/1168/2/022022.

[27] Gil Levi and Tal Hassner, "Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns."

[28] M. Liu, S. Li, S. Shan, and X. Chen. AU-aware deep networks for facial expression recognition. In Automatic Face and Gesture Recognition. IEEE, 2013

[29] Daniel Wolff, Tillman Weyde, and Andrew MacFarlane, "Culture-aware Music Recommendation."

[30] Shlok Gilda, Husain Zafar, Chintan Soni and Kshitija Waghurdekar(IEEE 2017) , Smart Music Player Integrating Facial Emotion Recognition and Music Mood Recommendation

[31] Ahlam Alrihaili, Alaa Alsaed, Kholood Albalawi and Liyakathunisa Syed, (IEEE 2019) , Music Recommender System for Users Based on Emotion Detection through Facial Features.

[32] Matthew A. Turk and Alex P. Pentland (IEEE 1991) , Face recognition using eigenfaces.