AUTOMATIC PREDICTION OF DEPRESSION AND ANXIETY

Gopika[1], Haritha H.[2], Sabira Reshni I.K. [3], Hema P. Menon[4]
Department of Computer Science and Engineering
Sreepathy Institute of Management & Technology,
Vavanoor, Kerala, India
[1]gopikap1999@gmail.com [2]hariharitha21@gmail.com
[3]ikreshni@gmail.coms [4]hemapmenon@simat.ac.in

Abstract — Mood swings, stress, anxiety these terms h0ave become more common in present day life. The persistence of these feelings in a person may lead to depression, anxiety, or both. Unfortunately, the majority of people's emotional / behavioural changes go unnoticed, which leads to a late diagnosis of the situation. This leads to a person's likelihood of harmful behaviors' including opioid, alcohol abuse or even suicidal tendencies. Hence, it becomes essential to identify such emotional changes in an individual. In this work a framework that uses a combination of attributes, including observable facial behaviour, modulations in the voice and self- reported personality ratings, has been developed to assess a person's mental health. The system also detects the gender of the person. The system uses Haar-Cascade classifier for detection of facial features and a Mel Frequency Cepstral Coefficient based voice analyzer. With this, an automatic prediction of anxiety, depression, or both can be achieved and appropriate support can be provided at the right time. The system has achieved a 70% of overall prediction accuracy.

Index Terms— Convolutional Neural Network, Harr-cascade classifier, Mel Frequency Cepstral Coefficient.

I. INTRODUCTION

"As the coronavirus pandemic rapidly spreads across the world, it has caused a considerable degree of fear, concern and anxiety among the entire population, especially in certain groups, such as the elderly, caregivers and people with basic health conditions". Many people are isolated, unable to interact with friends or family. Therefore, their mental health deteriorates. Some people use this time to improve themselves and find themselves learning and experimenting with new skills, while others are disconnected from everything and may slip into depression or anxiety. So a novel approach is proposed to detect the emotional state of a person to track their well-being. Depression and anxiety are frequently undetected and are responsible for a variety of morbidities, either directly or indirectly. While depression and anxiety are two distinct disorders, they often coexist. The main symptoms of depression from a clinical point of view are loss of memory, lack of concentration, an inability to make decisions, loss of interest in recreational activities and hobbies, overeating and weight gain or low appetite and weight loss, feelings of guilt, worthlessness, helplessness, restlessness and irritation, as well as suicidal thoughts. These symptoms were found to have a significant effect on important areas of an individual's life – such as in education, employment and social activities, and this provides a vital clue for forming a clinical diagnosis. The symptoms of Anxiety are irritability, nervousness, fatigue, insomnia, gastro-intestinal problems, panic, and a sense of impending danger, increased heart rate, sweating, rapid breathing and difficulty in concentrating. The proposed system analyzes a person's facial and speech emotional patterns to check whether they are depressed or anxious.

Depression is a mental health disorder characterized by a depressed mood, low self-esteem, remorse, and a lack of enthusiasm or interest in doing things. Anxiety is a natural and beneficial emotion. When an individual experiences disproportionate amounts of anxiety

on a regular basis, however, it can become a medical problem. Anxiety disorders are a set of mental illnesses marked by a high level of apprehension, anxiety, anticipation, and worry. According to the World Health Organization (WHO), about 300 million people worldwide suffer from depression, which is also one of the leading causes of disability. Almost two-thirds of patients with depression often suffer from comorbid anxiety disorders [5].

The traditional diagnosis of these disorders by medical practitioners takes a long time and is very costly. The majority of these are dependent on the patient's verbal responses. Technology progress has always resulted in improved living conditions. Recent studies have shown that machine learning and artificial intelligence can be used to monitor a person's mental health. The unconsciously transmitted behavioural symptoms expressed by head movements (pose), eye-gaze direction, and facial muscle movements (facial expressions) were used to model structures that could predict mental health conditions like depression and anxiety, and could be used as a valuable tool for clinical diagnosis of such conditions [1]. The relationship between personality and mental health problems such as depression and anxiety is a significant aspect that has gained little attention in the field of automatic behaviour understanding. Personality, which is characterized as a distinctive collection of behaviours, cognition, and emotional patterns, has been shown to be affected by genetic and environmental factors linked to mental illnesses including [6] depression and anxiety. Previous biological and psychological studies have frequently verified the existence of a significant connection between personality and mental health [7]. High neuroticism (one of the Big Five personality traits) has been linked to depression genetically [10][11][14], suggesting that neuroticism may represent a genetic vulnerability to depression. Other personality characteristics, such as extraversion and conscientiousness, have also been linked to depression [7]. This suggests that individuals with these personality traits are more likely to be depressed at some stage in their lives, and that certain personality traits can be seen as good predictors of depression.

II. LITERATURE SURVEY

With the latest advancements in the field of machine learning, especially automatic facial analysis, has generated much interest for the development of objective and repeatable methods to automatically analyze depression using behavioral data obtained from facial images. Previous approaches to acquiring human behaviour characteristics, action units, and various strategies for analysing depression levels are addressed in this chapter. Human behaviour primitives are automatically detected as low-dimensional multi-channel time series data, and two sequence descriptors are generated as a result. The first measures the behaviour primitive's sequence-level statistics, while the second transforms the problem into a Convolutional Neural Network problem based on a spectral representation of the multichannel behaviour signals [2]. Jaiswal et al. proposed a novel approach to Facial Action Unit detection based on a combination of Convolutional and Bi-directional Long Short-Term Memory Neural Networks (CNN-BLSTM), which jointly learns form, appearance, and dynamics in a deep learning manner [3], and an artificial intelligence system for monitoring depression is proposed, which can predict the Beck Depression Inventory (BDI-II) scales based on verbal and visual gestures [4] is discussed. Thus there has been a lot of research in this area.

Cohn et al.[5] looked into the connection between primitive human behaviour (Facial Action Units (AU), vocal behaviour) and depression. They came to the conclusion that this coding of human behaviour would provide crucial details for assessing depression. Primitive descriptors of human behaviour have a much lower dimensionality than video data, and systems for identifying them can be learned from non-clinical data. Two methods for extracting global depression features from human behaviour signals from a whole video was implemented by him. The first uses statistics to measure the results, while the second uses Convolutional Neural Networks (CNN). Since the human behaviour signals in each video are long and differ in length, and the CNNs need a fixed input dimensionality, these human behaviour signals cannot be used as the network's input directly. Several classification and regression experiments were performed on the DAIC-WOZ [6] database provided by the AVEC 2016 depression challenge to test the

efficiency of the proposed methods. The Facial Action Coding System (FACS), created by Ekman and Friesen, is a common standard for systematically categorising the physical expression of emotions by representing them as a combination of individual facial muscle actions (AU). The automated detection of facial Action Units is a challenging task because human facial behaviour differs from person to person and changes over time. Deep learning algorithms have been shown to enhance object detection accuracy for tasks involving object detection [11]. For facial AU recognition and facial expression detection, Jaiswal and Valstar used deep learning algorithms. The analysis of three facial features: face form, appearance, and dynamics is needed for good automatic facial AU recognition. These can be viewed as a source of additional data when it comes to the modelling of facial action unit detectors. According to Jaiswal and Valstar, learning all three features together will result in highly accurate models for facial AU recognition. However, how these are fused has a big impact on the models' results. They present a deep learning-based method for detecting facial AU in images in their paper. For AU detection, it employs Convolutional Neural Networks (CNNs) to model the appearance, form, and dynamics of facial regions.

III. PROPOSED SYSTEM

The proposed system architecture is shown in Figure. 1. It contains two modules: Voice Analyzer and Video Analyzer. The voice analyzer is used to detect and process the audio from the user in order to detect whether the person is depressed, anxious or not. And using the facial expressions that are observed, indicates the type of emotion the person is experiencing. The two are then integrated into a web page. From these two methods we are assessing and determining whether an individual is suffering from depression or anxiety. This system is an extension of the method proposed by S. Jaiswal, S. Song and M. Valstar in [1]. In this, a combination of observed facial behaviour and self reported personality information is used to predict depression and anxiety disorders using deep neural networks (DNN). The facial behaviour is encoded as a group of behavior primitives like facial Action Units (AUs), head-pose and eye-gaze direction. These frame or image level facial attributes are transformed into video level representation through either a histogram or spectral features. Variations in prosodic features can give more information into a person's mental health condition. Also both histograms and spectral features are used for representing the behavior attributes. This system aims to accurately detect both depression and anxiety and helps to warn a person before their condition becomes worse.

1) Voice Analyzer

The dataset provides 5 different emotions. They are calm, happy, sad, angry and fearful. Each audio file has a unique identifier in the sixth place of the file name that can be used to determine the emotion contained inside the audio file. To analyse and extract features from the audio recordings, Python's Librosa package is used. Librosa is a Python tool that analyses music and audio. It contains the components required to construct music information retrieval systems. Spectral features are extracted from the spectrogram. Spectrograms offer a powerful representation of the data. It depicts the power (dB) of a signal over time for a specific range of frequencies. This enables the system to identify recurring patterns over time and activity hotspots. Because they are concerned with detecting emotions from speech, the Audio Only zip file was chosen. Around 1500 audio files in wav format were contained in the zip file. The test was done on one of the audio files to know its features by plotting its waveform and spectrogram. Using the Librosa library MFCC features were extracted (Mel Frequency Cepstral Coefficient). MFCCs are a commonly used feature in autonomous speech and speaker identification systems. A periodogram is used to calculate the power spectrum of each frame, which is inspired by the human cochlea (an organ in the ear), which vibrates at different regions based on the frequency of incoming sounds. To do so, it is started by taking the Discrete Fourier Transform of the frame as shown in Equation 1.

$$S_i(k) = s_i(n)h(n)e^{-j2kn/N} \qquad (1)$$

where:

- $s_i(n)$ is the framed time signal (i frames)
- N is the number of samples in a Hamming Window
- $h(n)$ is the Hamming Window
- k is the length of the DFT

The identifiers can also be used to differentiate the female and male voices. Each audio file has a number of characteristics, which are essentially an array of several values. The labels are then added on top of these features. Then the missing features for some of the shorter audio recordings are fixed. The sampling frequency is increased twice to preserve the unique properties of each emotional language. The sample rate is not increased any further as it can collect noise and affect the results.

The next stage is to shuffle the data, divide it into train and test groups, and create a model to train the data. Convolution Neural Network appears to be the obvious choice given that the project is a classification problem. For our classification problem, the Conventional Neural Network (CNN) model performed best. The model uses an 18-layer CNN with 70% accuracy, softmax activation function, rmsprop activation function, 32 batch size, and 1000 epochs.. The summary of the CNN is shown in detail in Table I.

TABLE I
SUMMARY OF CNN

Layer (type)	**Output Shape**	**Param #**
conv1d 9 (Conv1D)	(None, 216, 256)	
activation 11 (Activation)	(None, 216, 256)	0
conv1d 10 (Conv1D)	(None, 216, 128)	163968
activation 12 (Activation)	(None, 216, 128)	0
dropout 4 (Dropout)	(None, 216, 128)	0
max pooling 1d 3 (Maxpooling1)	(None, 27, 128)	0
conv1d 11 (Conv1D)	(None, 27, 128)	82048
activation 13 (Activation)	(None, 27, 128)	0
conv1d 12 (Conv1D)	(None, 27, 128)	82048
activation 14 (Activation)	(None, 27, 128)	0
flatten 3 (Flatten)	(None, 3456)	0
dense 3 (Dense)	(None, 10)	34570
activation 15 (Activation)	(None, 10)	0

2) Video Analyzer

The video is taken from the webcam feed. Video is then converted into image frames. Faces are detected in each frame of the camera feed using the haar cascade method. The Haar-Cascade Classifier is a Machine Learning-based technique that involves training a cascade function from a large number of positive and negative images. Machine learning techniques are used in Haar Cascades to train a function from a large number of positive and negative images. This process in the algorithm is feature extraction. After that, it is used to find objects in other images. They are huge XML files. The system automatically detects faces using Haar-Cascade then crops it and resizes the image to a specific size and gives it to the model for prediction [8].

The first part of the Conventional Neural Network refers to three convolutional layers that can have spatial batch normalisation (SBN), dropout, and max-pooling in addition to the convolutional layer and Rectified Linear Unit (ReLU) nonlinearity, which is always present in these layers. The first convolutional layer has 64, 3x3 filters with a stride of size 2, as well as batch normalisation, dropout, and max pooling. It has 64, 3x3 filters with a stride of size 1, batch normalisation and dropout, and max-pooling with a filter size 2x2 in the second convolutional layer. In the third convolutional layer, 128, 3×3 filters, with the stride of size 2, coupled with batch normalisation and flatten and also max-pooling with a filter size 2×2 is present. The network leads to one fully connected layer after three convolution layers, which always has ReLU nonlinearity and can include batch normalisation (BN) and dropout. A hidden layer

in the fully connected layer is present with Softmax as the loss function and 512 neurons. ReLU is used as the activation function in all of the layers. Finally, the softmax loss function and scores are computed by the network. Seven probability values will be generated by the model, each corresponding to seven expressions. The highest probability value to the corresponding expression will be the predicted expression for that image. The region of image containing the face is resized to 48x48 and is passed as input to the CNN. The network generates a list of softmax scores for each of the seven emotion classes. On the screen, the emotion with the highest score is presented.

3) Web Page

For the implementation of the models, an open source web application is chosen. The purpose of this platform is to make available both emotion recognition models in an intuitive and easily accessible way. It allows users to obtain in real time a personalized assessment of their emotions or personality traits based on the analysis of a video and audio sent directly via the platform. The application has been conceived with the Python micro web framework Flask. Flask is a Python-based microweb framework. It is referred to as a microframework because it does not necessitate the usage of any specific tools or libraries. Each communication channel (audio, video) has its own web page, which allows the user to be evaluated. We start with an index page, which directs us to the facial detection and voice analyser. The facial detection page connects to the emotion recognizer, which evaluates the video's sentiment. We also added a PHQ-9 questionnaire so that the patient can answer these questions while being recorded. Then there's the voice analyzer, which records voices and detects the corresponding emotion from the recordings. The snapshot of the webpages are shown in Figure. 2 to 5.

In a video analyzer a VideoCapture(0) object is created to trigger the camera and read the first image/frame of the video. The function requests a frame from the camera, which it then returns as a response chunk with the content type image/jpeg. pyaudio.PyAudio() is used to record the audio from the user which is normalised and sent to a voice analyzer. Routes refer to URL patterns of an app. @app.route("/") is a Python decorator provided by Flask that allows us to quickly assign URLs in our app to functions. The decorator is telling our @app that whenever a user visits our app domain (http://127.0.0.1:5000/ for local servers) at the given .route(), execute the index() function. The Jinja template library is used by Flask to render templates. In our application, we use templates to render HTML which will display in the browser. The '/video' and '/voice' route returns the streaming response. The URL to this route appears in the "src" element of the image tag since this stream returns the images that will be shown on the web page. The browser will automatically keep the image element updated by displaying the stream of JPEG images in it, since multipart responses are supported in most browsers.

IV. DATASET

The datasets used to train the two CNNs are the RAVDESS dataset and the FER2013 dataset, the former for the Audio analyzer and the later for the Video analyzer.

A. RAVDESS

The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)is a multimodal database of emotional speech and song image that has been validated [12]. A database was constructed with a gender-balanced cast of 24 professional actors vocalising lexically matching statements in a neutral North American accent. There are two levels of emotional intensity for each expression, as well as a neutral expression. There are 12 male and female actors that record brief audio clips in 8 different emotions. Each audio file is named so that the seventh character corresponds to the various emotions. Each of the 7356 RAVDESS files has its own name. A seven-part numerical identifier makes up the filename (02- 01-06-01-02-01-12.mp4). The stimulus qualities are defined by these IDs. The file name identifiers are listed below

1) Modality (01 = full-AV, 02 = video-only, 03 = audio only).

2) Vocal channel (01 = speech, 02 = song).

3) Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad,

05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).

4) Emotional intensity (01 = normal, 02 = strong).

5) Statement (01 = “Kids are talking by the door”, 02 =
“Dogs are sitting by the door”).
6) Repetition (01 = 1st repetition, 02 = 2nd repetition).
7) Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).

B. FER2013

The data comprises 35,887 grayscale images of faces at a resolution of 48x48 pixels. The faces have been automatically registered such that they are more or less centred in each image and take up roughly the same amount of area [13]. The aim is to categorise each face into one of seven emotion categories, all of which are labelled, depending on the emotion expressed in the facial expression. The FER2013 dataset contains images that vary in viewpoint, lighting, and scale. Emotions in the dataset are ’Angry’, ’Disgust’, ’Fear’, ’Happy’, ’Sad’, ‘Surprise’ and ‘Neutral’.
train.csv contains two columns, “emotion” and “pixels”. The ”emotion” field contains a numeric code representing the emotion present in the image, ranging from 0 to 6, inclusive. For each image, the ”pixels” column has a string enclosed in quotes. The contents of this string are pixel values in row major order, separated by spaces. test.csv contains only the ”pixels” column. The training set consists of 28,709 examples. The public test set consists of 3,589 examples.

V. EXPERIMENTAL RESULTS

A. Voice Analyzer

The process of creating and tuning a model takes a longtime. The aim is to start simple and work your way up without adding too many layers merely to make it more complicated. With 18 layers, softmax activation function, rmsprop activation function, batch size of 32, and 1000 epochs, the CNN model achieved a validation accuracy of 70%. By integrating more audio samples for training, accuracy can be improved. It was also found that a long window of recorded audio affects the accuracy since the training audio set had 3 seconds of recorded audio, So the audio has been limited to 4 seconds. The accuracy of the training and validation set is given in Figure. 6. The real time output from various audio is shown in Figure 8.

B. Video Analyzer

The video Analyzer was excellent in detecting the number of faces in each frame. By default, this technology recognises emotions on all faces in the camera feed. With a simple 4-layer CNN, the test accuracy reached 63.2% in 50 epochs. The most common emotion is happiness, which has the greatest number of examples. Due to a large number of examples, Sad, Surprise, Neutral, and Anger are also good at detecting. Fear and Disgust perform worse, possible reasons would be Less training examples and for disgust, which is pretty similar to anger features. Sad emotions are also closely detected as neutral, because it's hard to distinguish them with just this much data. The accuracy of the training and validation set is given in Figure. 7.The real time output from various audio is shown in Figure 9.

VI. CONCLUSION

Nowadays, many people feel apprehensive and miserable every now and then because of depression and anxiety. In this work an automatic emotion prediction model was developed by using a combination of observed facial behaviour along with speech signal and self-reported personality information to detect depression and anxiety disorders. The incorporation of personality data improves the accuracy of the prediction. Unlike traditional methods, this strategy was able to get into the underlying link between mental health issues and personality features. The audio analysis module recognized five emotions as well as the person’s gender. The model was trained with the RAVDESS dataset. The model yielded an accuracy of 70%. The video analysis module recognized seven emotions in each frame, with the most common feeling being used to decide if the subject is depressed, nervous, or normal. The FER2013 dataset was used to train the CNN for facial analysis. The model yielded an accuracy of 63.2%. The voice and facial analyzer were combined using the flask framework, into a single web application to increase the overall efficiency. This framework would be useful in case of professional and medical interviews, to analyze a person’s current emotional state over a video call.
Such a system would help in providing timely assistance and thereby prevent occurrence of any risky incidents.

REFERENCES

[1] S. Jaiswal,S. Song,M. Valstar ”Automatic prediction of Depression and Anxiety from behaviour and personality attributes,” 8th International Conference on Affective Computing and Intelligent Interaction (ACII), 2019.

[2] Siyang Song1, Linlin Shen2, Michel Valsta, ”Human Behaviourbased Automatic Depression Analysis using Hand-crafted Statistics and Deep Learned Spectral Features,” 13th IEEE International Conference on Automatic Face Gesture Recognition, 2018.

[3] S. Jaiswal,M. Valstar ”Deep Learning the Dynamic Appearance and Shape of Facial Action Units,” In IEEE winter conference on applications of computer vision (WACV), pages 1–8. IEEE, 2016.

[4] Asim Jan, Hongying Meng,Yona Falinie A. Gaus, and Fan Zhang ”Artificial Intelligent System for Automatic Depression Level Analysis through Visual and Vocal Expressions,” , IEEE Transactions on Cognitive and Developmental Systems, 2017.

[5] J. M. Gorman, ”Comorbid depression and anxiety spectrum disorders Depression and anxiety,” 4(4):160–168, 1996.

[6] Min-Tzu Lo, David A Hinds, Joyce Y Tung, Carol Franz, Chun- Chieh Fan, Yunpeng Wang, Olav B Smeland, Andrew Schork, Dominic Holland, Karolina Kauppi, et al, ” Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders,” Nature genetics, 49(1):152, 2017.

[7] Yusuke Takahashi, Brent W Roberts, Shinji Yamagata, and Nobuhiko Kijima ”Personality traits show differential relations with anxiety and depression in a nonclinical sample” Psychologia, 58(1):15–26, 2015.

[8] Padilla, Rafael Filho, Cicero Costa, Marly. ”Evaluation of Haar Cascade Classifiers for Face Detection”. 2012

[9] Li, Qin Yang, Yuze Lan, Tianxiang Zhu, Huifeng Wei,
Qi Qiao, Fei Liu, Xinjun Yang, Huazhong.” MSP-MFCC:
Energy-Efficient MFCC Feature Extraction Method with Mixed-Signal Processing Architecture for Wearable Speech Recognition Applications”, IEEE Access. PP. 1-1. 10.1109/ACCESS.2020.2979799, 2020

[10] John M Hettema, Michael C Neale, John M Myers, Carol A Prescott, and Kenneth S Kendler, ”A population-based twin study of the relationship between neuroticism and internalizing disorders,” American journal of Psychiatry, 163(5):857–864, 2006.

[11] Kenneth S Kendler, Margaret Gatz, Charles O Gardner, and Nancy L Pedersen, ”Personality and major depression: a swedish longitudinal,” population-based twin study, Archives of general psychiatry, 63(10):1113–1120, 2006.

[12] Livingstone SR, Russo FA (2018) ” The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English.” PLoS ONE 13(5): e0196391.

[13] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, ”Challenges
in representation learning: A report on three machine learning contests. Neural Networks”, 64:59–63, 2015. Special Issue on ”Deep Learning of Representations”

[14] Nemesure, M.D., Heinz, M.V., Huang, R. *et al.* Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. *Sci Rep* 11, 1980 (2021).

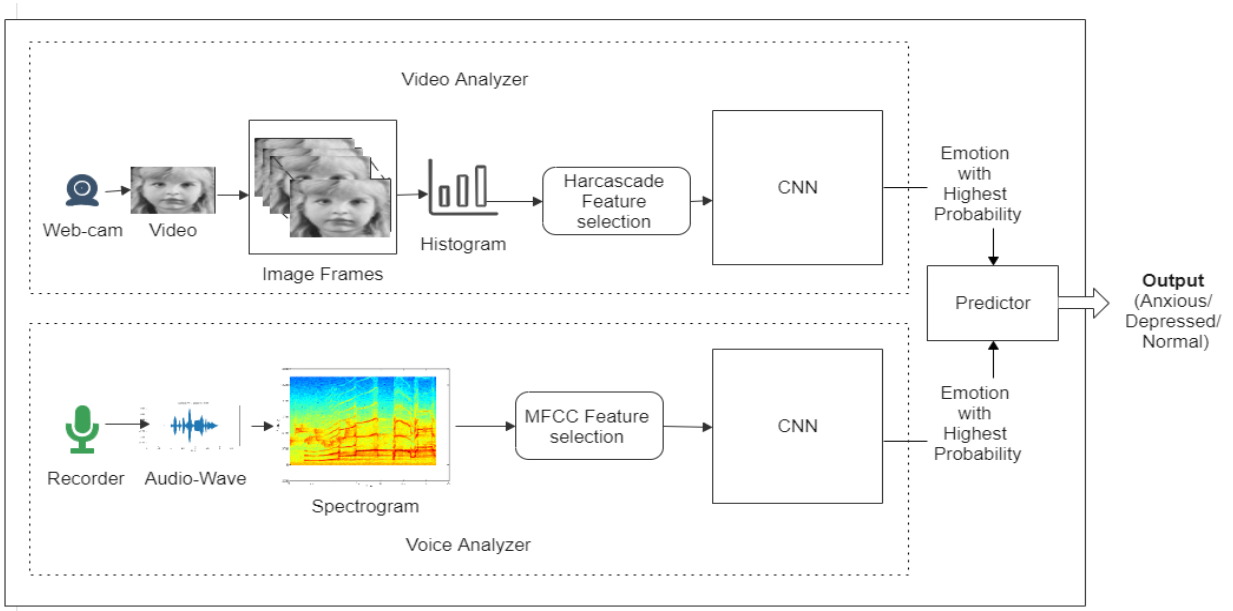


Figure. 1 System Architecture. There are two emotion analyzers- video analyzer and voice analyzer, the two modules use separate CNNs and produce an output with highest probability. That is Anxiou, Depressed and Normal.

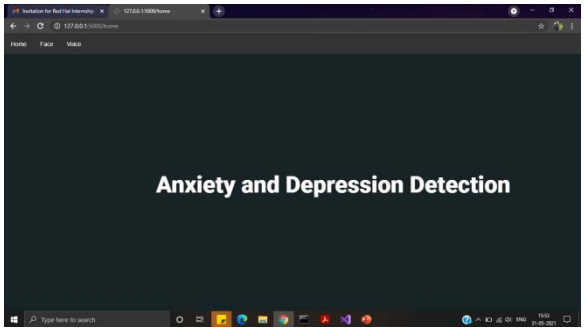


Figure 2.Index Page

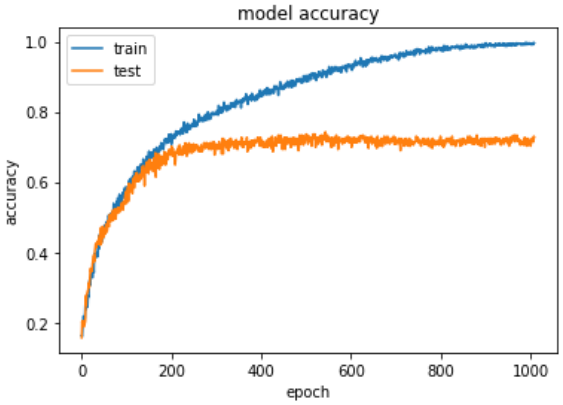


Figure 3. Voice Analyzer

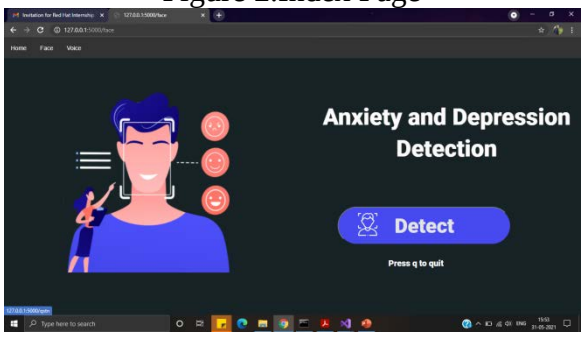
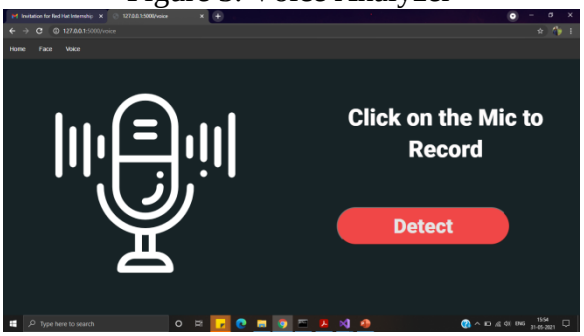


Figure 4.Video Analyzer



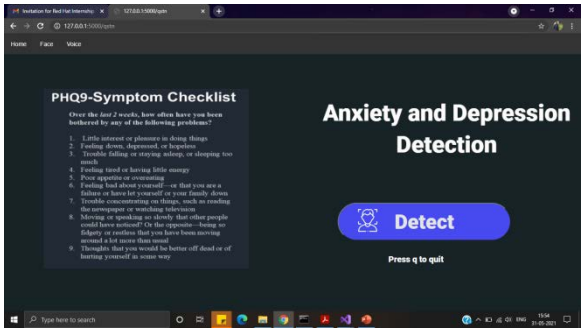


Figure 5. PHQ9 Questions

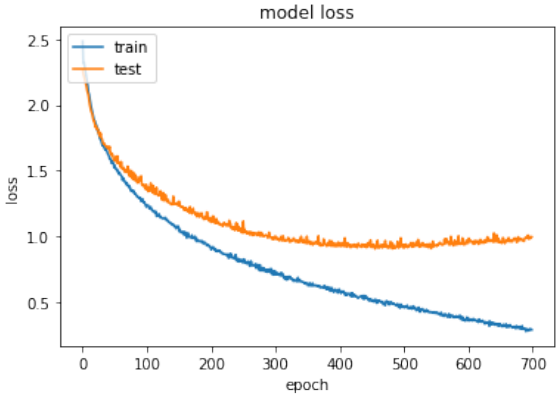


Figure 6. Model Loss and Model Accuracy on test and train dataset of voice analyzer.

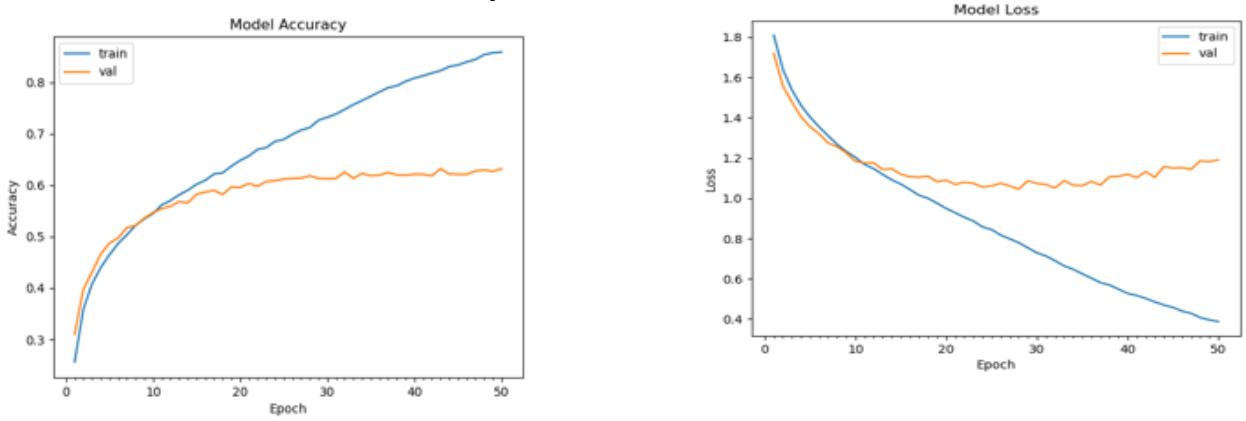


Figure 7. Model Loss and Model Accuracy on test and train dataset of video analyzer.

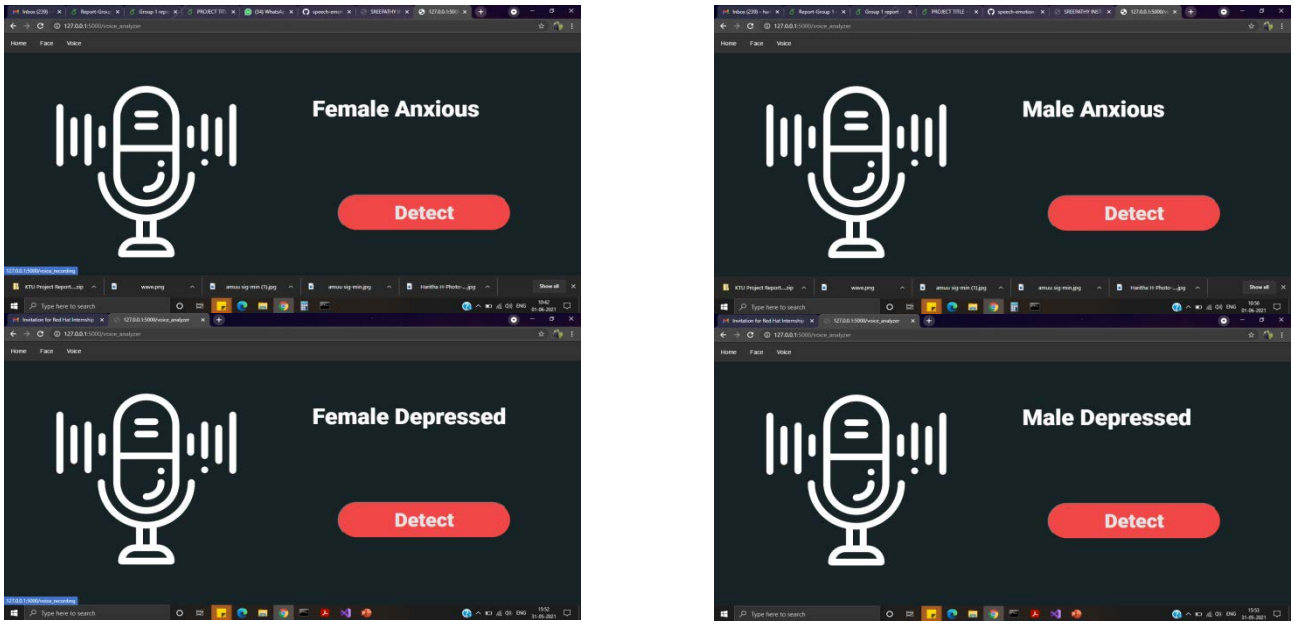
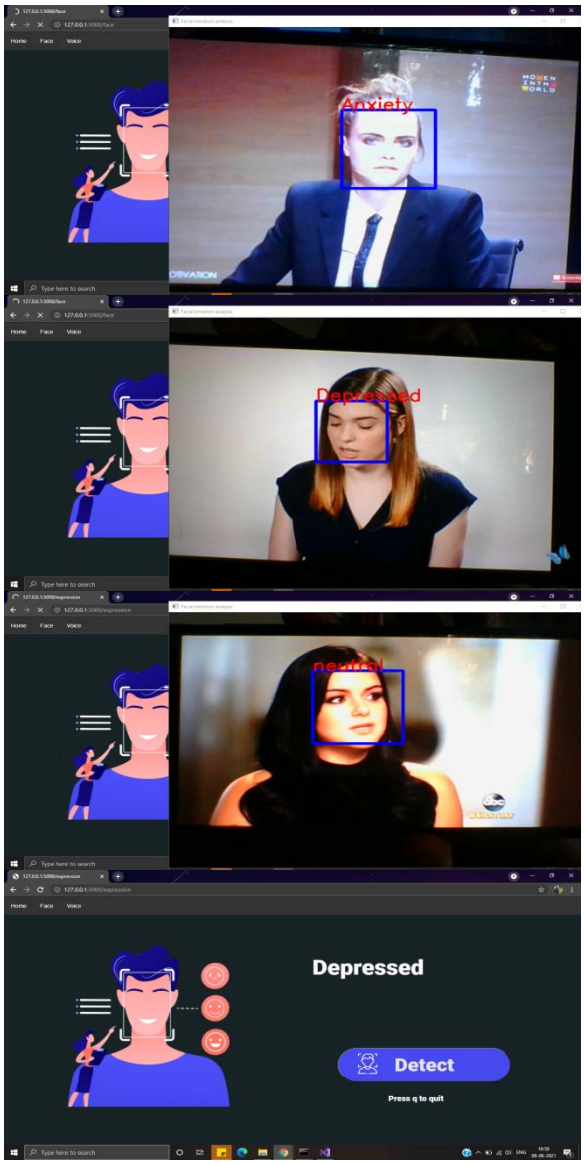


Figure 8. Output from the voice analyzer



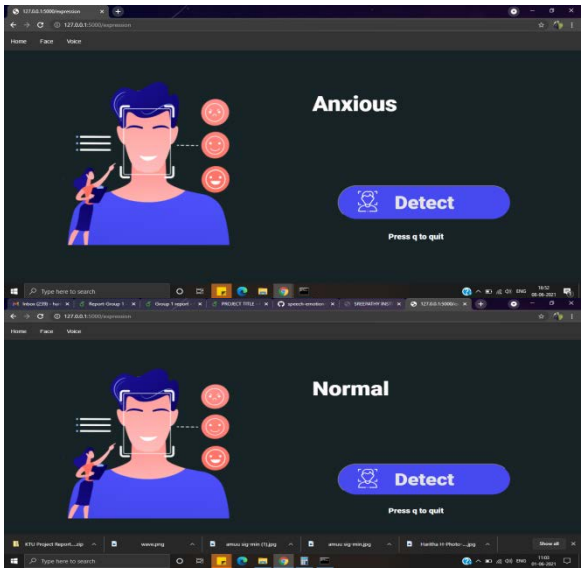


Figure 9. Output from the Video Analyzer.