Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5341 - 5349
Received 15 December 2020; Accepted 05 January 2021.

Depression Detection from Facial Behaviour through Deep Learning

Abhijit Biswas[1] , P Sandhya[2]T.R.Saravanan[3]
[12]School of Computer Science & Engineering, Vellore Institute of Technology, Chennai, India
[3]SRM Istitute of Science and Technology,Kattankulathur,Chennai,India
Email- abhijitbiswas1999@gmail.com[1],sandhya.p@vit.ac.in[2]saravanatrcse@gmail.com[3]

Abstract. Millions of people worldwide suffer from depression. There are some differences in condition of mental health between two people who have the same disorder. The degree of depression is analyzed through video-recorded clinical meetings. Depression is a mood disorder and is a challenging issue in now days. In worldwide there are 350 million people suffering from depression. Depression patients find hard to concentrate on their work. They may suffer from insomnia, restlessness, loss of appetite and sometimesthey also get suicidal thoughts. In this paper we discuss about depression detection methods. This paper elaborates how Hand-Crafted Method, Deep Convolutional Neural Network, Raw Audio, Spectrogram Audio, Local Binary Pattern, Median Robust Extended Local Binary Pattern, Joint Tuning, Naïve Bayes Algorithm, Gaussian Mixture Model etc. are used in depression detection. Here we discuss some audio, video and text recognition methods and try to find best detection technique for depression.

Keywords: Depression Detection, Deep Convolutional Neural Network, Naïve Bayes Algorithm

1. Introduction

Depression is a mood disorder and is a major issue in now days. In this paper we propose a method to examine patients through computer vision for depression detection. Depression detection can be done through audio visual recording. The depression is detected by analyzing the victim's eye gaze, eye contact and speaking. The voice patterns are divided into major parts like prosodics, cepstral, spectral, vocal tract and glottal source[1],[15-16]. The above parameterscan be used to analyze the patients and give psychotherapy.In this paper we mainly use Beck Depression Inventory (BDI) depression detection technique for better performance. Nowadays Hamilton Rating Scale is used as the standard by clinicians to determine the degree of depression. There are many other techniques at present like Quick Inventory Symptoms Self Report (QIDS). Here we mainly use BDI technique for detecting the power of the depression.

In recent years some machine learning methods have been proposed for detecting the causes of depression and the intensity of depression which can be used to arrive at a treatment strategy for depression. Some researchers say that we can understand or guess that any person is depressed or suffering from any problem or not from his or her voice. In this paper we use voice as a parameter in depression detection. We record patient's voice, face gesture, eye gaze and head position. There is certain Hand crafted features can be used to predict depression severity but has many limitations. The main limitations are hand crafted featuresneeds lots of efforts to design such type of system. If we design hand crafted features like MFCCs, then it is necessary to have good knowledge about it which is time-consuming. Secondly it may result in loss many useful information related to depression patterns.

Recently Deep learning method has been used for depression detection. In this paper we choose to use deep learning fordepression detection. In deep learning features CNN means convolutional neural network has been widely used to achieve its best level of work. We use many data sets like AVEC13, AVEC14, AVEC16. In this work in use raw DCNN (Deep Convolutional Neural Network) and spectrogram DCNN and then analyze them and convert them into hand crafted feature and MRELBP (Median Robust Extended Local Binary Pattern) respectively. We also use BDI where some set of questions are asked to the patient and a video recording is taken and analyzed. Depending on the patient's answer a probabilistic result is generated and the severity level is determined. Depending in

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5341 - 5349
Received 15 December 2020; Accepted 05 January 2021.

this intensity suggestions and treatmentare given to patient.

The main aim of this work is to regularly go for checkup for extra cure. This work enables to detect the depression in early stage and notify the patient very early and patient can try to stay depression free. We have useddeep learning algorithms so it gives very accurate result. It is very useful for clinicians to use this computerized system for detection of depression. We have used many datasets for give accurate result for patients.

2. Literature Survey

In this section we discuss about some previous researches in depression detection. This section is divided into some parts like audio feature and video feature.

2.1. Audio Features

Combine hand crafted features and deep learning features for better result and increase the power of the detection technique. At first, they get low level descriptors from raw audio and median robust extended local binary patterns from spectrogram audio samples for hand crafted features. After that we use DCNN to easily adopt the knowledge about the deep learned features from the audio's like raw audio and spectrogram audio. At last they combined the two features like hand crafted features and deep learned features and called this combination as joint tuning technique or method for better result.

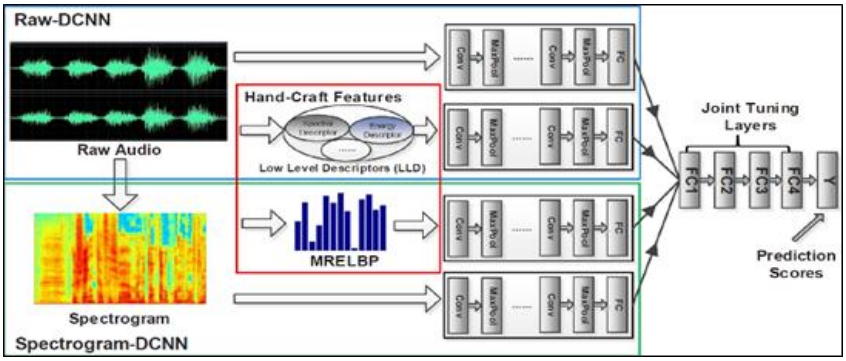

Figure1. Hand crafted based feature extraction [1]

In the given figure1 they try to show that the raw DCNN takes the raw audio as input and low-level descriptors as input where spectrogram DCNN uses features of texture as input in this method. At last the result is given based upon the average of prediction of one by one per frame from 4 deep convolutional neural network.

Then MRELBP is used to calculate LBP feature to detect the local structure of the image of the spectrogram for further analyze. It is performed by encrypting the differences between the value of the pixel that is present in the middle and to those who are neighbors of this pixel, finally it gives a pattern in binary. Generally we assign a variable *Xc* in the image that is the center or middle point through computing its value of *P* means neighbor pixels $\{X_R, P, n\}$, The radius of the circle point is calculated as *Xc*-

$$LBP_{R,P}(X_c) = \sum_{n=0}^{P-1} (X_{R,P,n} - X_c)$$

$$S(X) = \begin{cases} 1, \text{if } X>=0 \\ 0, \text{if } X<0 \end{cases}$$

The above formula is used for calculate the binary pattern of the image of spectrogram.

After that they discuss about deep learned texture features. Here they describe the details for the deep learned texture features. It is another form of deep learned feature. Few years ago, CNN was used but is not the best solution because ithad many disadvantages. CNN can't store high resolution spectrogram images; another disadvantage is it needs large number of training of samples, datasets training. It is very time consuming and also not cost efficient. To remove the limitations they used

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5341 - 5349
Received 15 December 2020; Accepted 05 January 2021.

deep learned texture features. By which they split audio into segments of 6 seconds to 20 seconds. Then they use data augmentation technique to do work with the small samples of the audio training data samples. First some features are extracted then the whole spectrogram images are flipped to horizontal image. Every image is rotated by 6 angles that are -15^0 , -10^0 , -5^0 ,5^0,10^0,15^0. Finally they got 14 times more data than the previous original image. Like original image, rotated image, flipped image etc. The picture frame used here is 5x5 for better result previously they used 128x128. However if they use 128x128 image ratio then the result is not perfect because the final data will out have bound to the frame. So, they use 5x5 ratio of the spectrogram image and after that they got best results. So, they now use this ratio.

In [2] PrathameshRaut et al discusses, few years back that many depression detection works are done only through BDI. Now days some researches uses some BDI with some audio features.For example from BDI some questions are asked to you write your answers and also you have tell your feelings like what you are feel now or how you feel at some specific time.This is captured via a microphone and your voice will be recorded after which your voice can be analyzed.By the help of googleapiit helps to convert the voice to a text and then analyze it. Finally a decision can be arrived based on the analysis.

In [3] ShubhamDham et al discusses,in AVEC datasets there exists many audio features or audio data that is used in their work for efficient result. These are collected from interview audio data(s) that were used in their work. The data(s) are pre-extracted from COVARREP toolbox at 10-ms intervals over the entire interview. They present a concept of transcript file which contains speaking times and values of patient and some virtual interviews etc. The given audio file was processed so as to obtain the voice of the participant only. Hence from the audio file, the voice of the participant was isolated using the speaking times given in transcript file. From this audio two sets of functionalities or features are calculated. That sets are given in the bellow table1-

Table1. Statistical descriptors calculated from the pre-extracted audio features [3].

Low Level Description	Statistical Features
Normalized F0, NAQ, QOQ, H1H2, PSP, MDQ, peak Slope, Rd, Rd conf, MCEP 0-24, HMPDM 1-24, HMPDD 1-12	Mean, min, skewness, kurtosis, standard deviation, median, peak-magnitude to root-mean-square ratio, root mean square level, interquartile range

All the features are calculated by MATLAB. The LLDs are provided in the covarep file with data sets. And the other set of features are concise with DCT (Discrete Cosine Transform) coefficients for each descriptor in the first column of the table. At last the two sets of data are combined and analyze further.

2.2. Video Features

In [1] Lang He et al discusses that,AVDLC subsets are existing in AVEC13 dataset. It consists over 340 videos and of over 292 subjects, performing a recording of human and computer interaction with a webcam and a microphone. It consisting of some different power point guided task like 1-10 number counting, sustained smiling vowel phonation, sustained vowel phonation, sustained loud vowel phonation and speaking out loud while solving any kind of task. These subjects are one to four-time recorded for better prediction in two weeks of time period. The age of the persons is 31-year-old. Each video is 20 minutes to 50-minute-long with the audio sampling rate 16bit and also videos are recorded with 480p with 24bit sampling rate ,30 frames per second for better analysis and after that that all are divided into 3 parts that are training, developing and testing.

After all of these are calculated the depression level was detected through BDI-II dataset's score. The score is between 0-63 means total 64 levels of score were there. And the level is divided for power of depression-

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5341 - 5349
Received 15 December 2020; Accepted 05 January 2021.

Table 2. BDI level wise power analysis [1].

Level	Power
0-13	Low
14-19	Normal
20-28	Powerful
29-63	Very Powerful

The corpus of AVEC14 is a subset of AVEC13 corpus. The extra feature is in AVEC14 is, it's had 2 different human computer interaction dataset tasks. That 2 tasks are delivered from two different recordings. So, it's requires greater number of videos and subjects. Here we have over 300 videos but they are not as long as AVEC13. Here the time duration of the videos are 4 to 6 minutes. There they use two types of methodology, that are-

2.2.1. Northwind

Here patients read aloud a collection of the tale "Die Sonne und der Wind".

2.2.2. Freeform

Here patients respond to some numbers of questions like- "What you like most", "What is your favorite things" or "Say about your childhood memory good or bad anything you want to share".

After collecting all of these again like AVEC13 are divided into 3 parts as discussed before i.e. train, develop and last of all test.

In [2] PrathameshRaut et al, used digital image processing and MATLAB for video processing and successful recognition of facial features of a patient. In this system the patienthas to be faced towards the camera and then the facial expression of the patient is captured through the camera. Then the above-mentioned toolslike digital image processing and MATLAB constantly track the lips, eye gaze and head of the patient. It concentrates on patient's lips because when you are depressed thelips might be not look as usual. So, it detects the lips and continuously keep track on it.

In [3] Shubham Dham et al mainly focused on video features for depression detection. Because it is really very secure way to detect through video. Here you clearly see patients face head pose, eyes and all of his activities. Mostly physicians directly face the patients and analyze if they are depressed. It might be fail sometimes. This paper discusses about head pose or head features. Here motion of patient's head track by horizontal and vertical motion and of some facial points is analyzed. Like 2,4,14,16 points are used to track the vertical and horizontal position of the face as you can see in figure 2. The points are engaged with a person's smile, cry, blinking and other facial expressions. Those points are representingevery changes of motion of patient's head. Every point is assigned for different jobs. If any position is changed then the change in their position are calculated between all neighbor points. Every change is measured as horizontal and vertical change and measured in some statistical methods like mean, median, mode of changes points in horizontal and vertical directions and magnitude of this change, velocity in horizontal, vertical direction and magnitude of velocity. All the data are sampled as change between two neighbor frames that should be ignored.

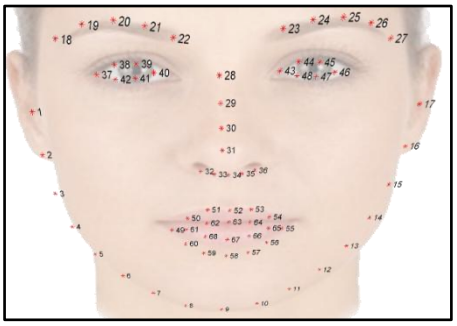

Figure 2. Facial Pointing for tracking

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5341 - 5349
Received 15 December 2020; Accepted 05 January 2021.

Here videos are recorded and then checkedfor each and every points and then calculated all of these after which the result is generated. The distance betweenpoint {37,40} was used to track the left eye's horizontal distance like that for right eye these points are {43,46} and for left eye the vertical points are {38,42} and for right eye {44,48}. For mouth the horizontal points are not only one point their have two points pair that are {55,49}{65,61} and the vertical points are {55,58} for the head the horizontal points are average of {2,16} and {4,14} and vertical also average of two points {22,8} and {23,10} there also exist eyebrows that simultaneously calculated through point {22,23} {27,18} pairs of points for horizontal. For nose tip point pair are {31,25} and {31,20} used for vertically tracking the nose and calculate the points. The distance was calculated for each and every frame and it is giving result as 10 vector per video. This vector is very useful because the shape of every face is different, we here done 10 vector's some and then a value is generated that use for different shapes of face and also help to detect the characteristics of the face.
After all of these vectors are processed by GMM (Gaussian Mixture Model) it generates trunk of words. Then the important vectors are extracted for each and every video from the cluster using GMM. It presents the probability of points associated to a cluster. According to this model the probability is distributed almost the mean of the cluster. The resultant clusters are after divided into total of 64 clusters using k-mean method. Then the resultant clusters are used as an input of EM (Expectation Maximization) for making GMM.

Then the eye blink rate is calculated using a 2d representation of the face. The point taken for this is {37-42}. For these points a polygon is generated and then the data in eye area VS frame data was gained. Now to calculate the opened eye area the mode of total numbers of area per frame is calculated. This time they take is minimum 1000 of random frames. A blink eye is considered if 90% area of an eye is closed. The blink frequency was calculated in this by counting the blink over the interview. The bellow figure 3. shows the comparison between them.

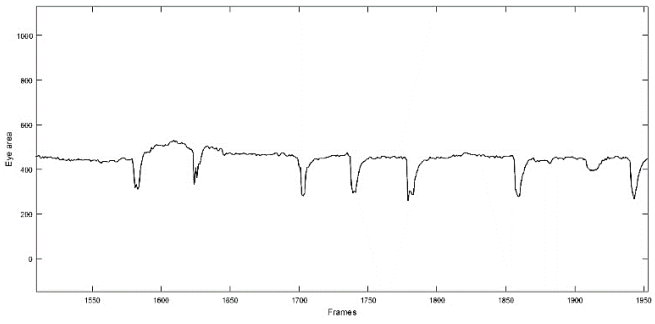

Figure 3.Eye area VS frame data [3]

For visual features the history of motion images are needed. Hence the motion images are converted into grayscale images, and it consists of information about the most recent movement or change of the pixel with the highest movement value.

Here mainly two classifiers are used one is SVM and the other is neural network. SVM classifier was applied for all the extracted features to perform its task. Here 8 models were trained. The models are PHQ8_nointeresr, PHQ8_depressed, PHQ8_sleep, PHQ8_tired, PHQ8_appetite, PHQ8_failure, PHQ8_ concentrating, PHQ8_moving. At last all the 8 model are given score 0,1,2 or 3 and all are added for total score out of 24. Neural network is applied only to the fisher calculated. Because the total dimension of other features is very small and these small dimensions are not useful for understanding or make sense to anyone. Hence this approach uses neural network that provides much better results. Here also another type of neural network that is regression neural network was also trained for giving the combined result of PHQ8.

3. Proposed Method

Behavior of a depressed person disturbs himself and others around him as well. So, it is good to take a proper checkup or examination from a good psychological expert and take counselling for better health. Depression is just a change of mental health and state.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5341 - 5349
Received 15 December 2020; Accepted 05 January 2021.

In this work we are going to discuss about our work on depression detection through deep learning. Here in this work we are going to take a picture of a patient through webcam and save those pictures in JPEG format and use those pictures for analyzing the depression. For this we have to make a model for testing the picture whether the picture is of a depressed person or a normal person. For that, at first, we have to train our model. Training of a model means make the system capable to predict something, adapt, and judge something. To achieve this we have to train the system or machine or our model. In our work we have used some deep learning approaches. Here we use CNN (Convolutional Neural Networks) for training data. We must need dataset for making the network capable of predicting the result. So, we use dataset downloaded from Kaggle named "facial expression recognition challenge (fer2013.csv)" [9]. In this dataset we have emotions and their corresponding gray scale pixel values in a 48X48 matrix form, we have test data and train data. Train data is used for training our CNN model to predict something, and test data for testing our model[10][11]. For training data, we have almost 80% of data and for testing we have almost 20% of data[9]. We are not saying that it is the best way to train and test any model. The best or standard way to train any model is with 70% data and test the trained model is with 30% data.

After that we are going to discuss about training our model in details. So, for training the model we have usedPython. In python we have many mathematical libraries that are very useful to calculate very difficult problem in a very short time. So, we use that libraries for my code to make this code very sort and effective[6]. We have used "KERAS"[6] that is very famous for deep learning means making deep learning networks, after that we use "NUMPY" , for working with multi-dimension array, then we use "OPENCV"[8] for working with images, like taking real time images, process those images for analysis. The main thing of making network we need huge amount of processing power, but for our local laptop or desktop it is very difficult task to process that much data at a time. If we do so it takes huge amount of time and waiting for that time, we will be frustrated some time the machine doesn't work properly. So, we use "google colab" for running my code. It gives us unlimited processing power, unlimited storage and so on. In colab we can use GPU (Graphical Processing Unit) it can do our model or network making task very fast than CPU. For our model we use 256 as batch size and epoch as 25, it is not a fixed size, it can be any integer number as your choice, larger the epochs the better model is. A batch size is like a for-loop iterator for one or more samples of data and making prediction after that a result is generated then the result is compared with expected result, and an epoch means each sample data have an opportunity to upgrade itself for further iteration, and one epoch is,consist of many batches for training the neural network model[12]. After that we make convolution layers, here we make 3 convolution layer and 1 fully connected layer that gives the final output layer and it looks like that-

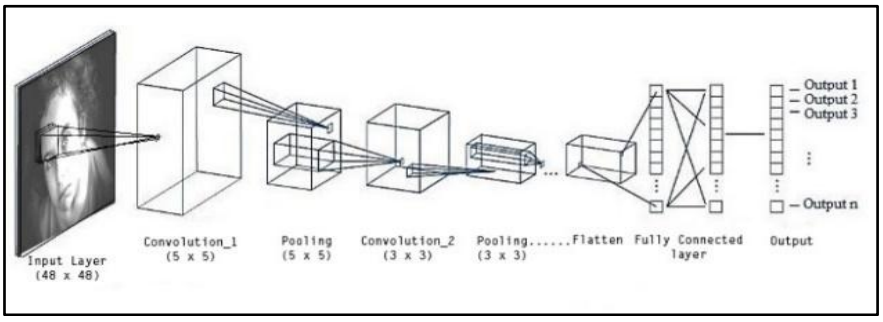

Figure 4. Input to output or Neural Network Modeling Process

In our model we use activation function RELU (Rectified Linear Activation Unit)[11]. An activation function is used to make the input weights into activated node or the output. There are many activation functions are available but we use relu because it works as linear activation function sometimes when the inputs are in positive otherwise 0. It uses the formula- ***f(x)=max (0, x)***

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5341 - 5349
Received 15 December 2020; Accepted 05 January 2021.

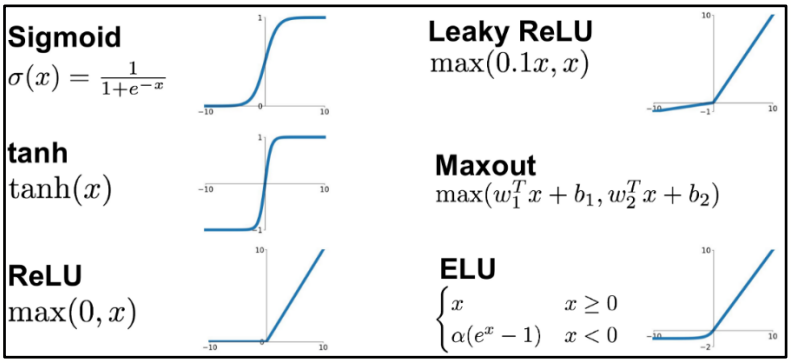

Figure 5.Curves of some of the activation function [7]

We use this because it looks like a linear activation function that is very easy to understand and do any calculation but a drawback of a linear activation function is it is not capable to do very complex work or calculations, for doing that type of calculations we need a non-linear function that work is done by "RELU" here [11]. As mentioned before it is works like linear as well as non-linear activation function. After that we test model for its accuracy and loss. In the below diagram the model's accuracy and loss are plotted. We found our model's accuracy as 96.57% and loss as 9.85%, the figures are given to the next page-

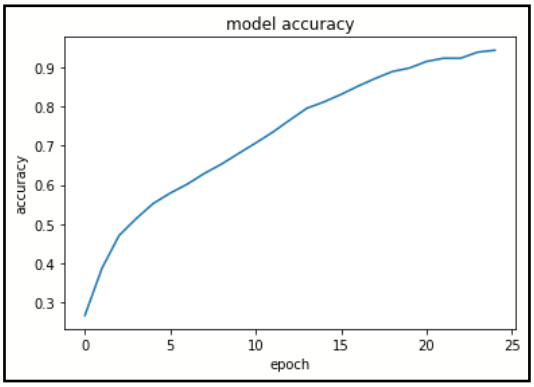

Figure 6. Model Accuracy curve

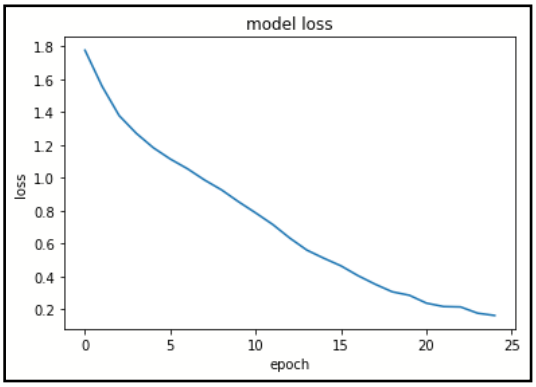

Figure 7. Model Loss Curve

The real-time image of patient or who came for checkup is taken and "OPENCV"[8] library is used for image pre-processing. We have taken images through camera and saved them, read any picture from any storage and perform some image pre-processing tasks[8]. So, we are using this OPENCV library in python to crop the facial area into a 48X48 pixel image. The next task of analyze the picture or the input become easy for the model. Then we read that image and load to the model. Before loading the image, we make the image as grayscale image because grayscale image is very powerful for any image pre-processing task. After that we pass the 48X48 image through the model. The model will give us the result as a bar chart consist of moods of a human being in percentage. From that the depression status of a patient can be determined. The results are given bellow-

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 1, 2021, Pages. 5341 - 5349
Received 15 December 2020; Accepted 05 January 2021.

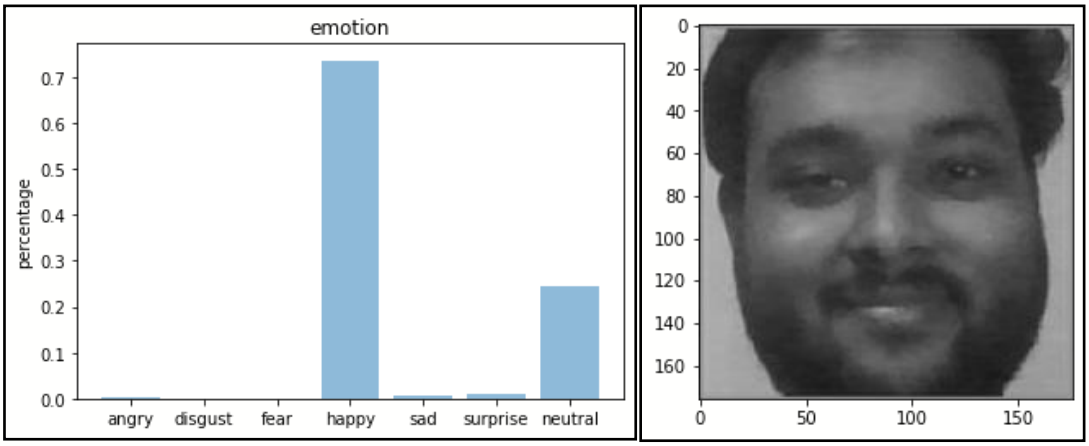

Figure 8.Happy face Bar & Input Image

Figure 9. Natural Face Bar & Input Picture

As you can see above in figure 8 the face is happy so the bar goes high for happy approximately 70% and about 25% natural. For the face in figure 9 it predicts and classifies as 100% natural because it is not depressed and shows natural.

4. Conclusion and Future Work

In this paper we discussed some of the technique that helps to detect depression. There are many techniques discussed in the three papers but have certain limitations. The first approach is good to detect through voice but not well for video and then third paper is vice-versa. So we can adopt the audio extraction features from the paper [1] and video extraction feature from the [3] along with our approach for depression detection through facial behavior. We have also analyzed the results of our work.

5. References

[1] J.Sangeetha,T.Jayasankar,“ A Novel Whispered Speaker Identification System Based on Extreme Learning Machine”, International Journal of Speech Technology, Springer,(2018) 21 (1), pp.157–165. DOI:https://doi.org/10.1007/s10772-017-9488-z

[2] PrathameshRaut , Pratik Kalbhor , HuzefaHirani , LaveenRaheja , Prof. P. Y. Pawar, “Depression Detection using BDI, Speech Recognition and Facial Recognition,” International Journal for research in Applied science and Engineering Technology Vol.6 ,No.4 , pp.347-351, April 2018

[3] ShubhamDham , Anirudh Sharma , AbhinavDhall , “Depression Scale Recognition from Audio, Visual and Text Analysis,” Computer Vision and Pattern Recognition , Cornell university , https://arxiv.org/pdf/1709.05865.pdf , Sep 2017

[4] Xavier Glorot, Antoine Bordes , YoshuaBengio “Deep Sparse Rectifier Neural Networks,” n

Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS) 2011
[5] David Sussillo, L. F. Abbott , "Random Walk Initialization for Training Very Deep Feedforward Networks," arXiv: Neural and Evolutionary Computing,pp.1412-1422 2014
[6] Hasbi Ash Shiddieqy, FarkhadIhsanHariadi, Trio Adiono ,"Implementation of deep-learning based image classification on single board computer," International Symposium on Electronics and Smart Devices (ISESD),17-19 Oct. 2017
[7]Ochin Sharma "A New Activation Function for Deep Neural Network," International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon) ,14-16 Feb. 2019
[8] MauricioMarengoni, Denise Stringhini ," High Level Computer Vision Using OpenCV," 24th SIBGRAPI Conference on Graphics, Patterns, and Images Tutorials,28-30 Aug. 2011
[9] JyotiKumariR,RajeshK,M.Pooja , "Facial Expression Recognition: A Survey," Procedia Computer Science - Elseiver ,Vol.58,pp486-491,2015
[10] Jianli Feng, Shengnan Lu ,"Performance Analysis of Various ActivationFunctions in Artificial Neural Networks ," Journal of Physics: Conference Series 1237 022030, pp1-6,2019
[11] Sangeetha .J, Jayasankar.T , "Emotion Speech Recognition based on Adaptive Fractional Deep belief Network and Reinforcement Learning, Springer- Advances in Intelligent Systems and Computing - International Conference on Cognitive Informatics & Soft Computing (CISC-2017),pp-167-174.https://doi.org/ 10.1007/978-981-13-0617-4_16
[12]HoneyJindala, NeetuSardanab , RaghavMehtac , "Analyzing Performance of Deep Learning Techniques for Web Navigation Prediction ," International Conference on Computational I Intelligence and Data Science (ICCIDS 2019) ,Science Direct, Vol.167,pp.1739-1748,2020
[13]YatharthYadav, Vikas Kumar, VipinRanga, Ram MurtiRawat ,"Analysis of Facial Sentiments: A deep-learning Way," International Conference on Electronics and Sustainable Communication Systems (ICESC) , 2-4 July 2020
[15]P.Shanmugapriya, V.Mohan, T.Jayasankar, Y.Venkataramani, "Deep Neural Network based Speaker Verification System using Features from Glottal Activity Regions", Appl. Math. Inf. Sci. vol.12, no.6, Nov 2018, pp.1147–1155.
DOI: http://dx.doi.org/10.18576/amis/120609
[16]Donepudi Babitha, Jayasankar.T, Sriram V.P, Sudhakar S, Kolla Bhanu Prakash," Speech Emotion Recognition using State-of-Art Learning Algorithms", International Journal of Advanced Trends in Computer Science and Engineering, Vol. 9, No. 2, pp.1340-1345,March-April (2020).
https://doi.org/10.30534/ijatcse/2020/67922020