2022 6th International Conference on Electronics, Communication and Aerospace Technology (ICECA) | 978-1-6654-8271-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICECA55336.2022.10009064

Deep Learning based Audio Processing Speech Emotion Detection

Dr M Kavitha
Department of CSE
Koneru Lakshmaiah Education
Foundation, Vaddeswaram, AP, India
mkavita@kluniversity.in
http://orcid.org/0000-0003-1963-8330

B Sasivardhan
Department of CSE
Koneru Lakshmaiah Education
Foundation, Vaddeswaram, AP, India
sasivardhanb@gmail.com

P Mani Deepak
Department of CSE
Koneru Lakshmaiah Education
Foundation, Vaddeswaram, AP, India
deepakchinnuaug3@gmail.com

M Kalyani
Department of Mathematics
PACE institute of technology and
sciences, Ongole, AP, India
Modepallikalyani123@gmail.com

***Abstract*— This paper represents the usage of deep learning model to differentiate the audio emotion classification from a given speech, basically it is called as Speech Emotion Recognition (SER). The intensity of voice is useful to determine the pitch and tone of voice helps to differentiate the emotions based on the audio. In this paper the Multilayer Perceptron model helps in classifying the emotions from the audio. RAVDESS emotional audio speech dataset is used in this work. Feature selection methods are applied to select required features. The data set is more comfortable for extracting features. Model is trained using extracted features and the predicted results are verified based on accuracy parameter. Web application is designed to access the results at user end.**

Keywords—Deep learning, Emotion detection, Feature extraction, MLP Classifier

I. INTRODUCTION

Speech is one of the important feature for the human communication and also a mainly it is one of the important human behaviours. Speech Recognition is one of the concept in the domain of artificial intelligence where we can apply this on the systems to identify the speech information . Our main idea is to find the emotion and accuracy of correctness from the audio that is going to be input and representation of the outcome in the application format [1].

It is one of the trending and important topic in the current technical world. Emotion is a path that a person expresses his feeling [6-7]. Emotions play an vital role in every domain in the it sector and in the army field and also in the health care sections one need to keep finding it . Finding and prediction of emotions is literally a difficult test because there are different type of emotions for different persons and they perform their emotions in their own tone [8] . various types of emotions that are commonly seen are happy , neutral , angry , sad .etc... So to differentiate different type of emotions by a proper technical method and representing the outcomes of this work in the form of web application and we used flask for this web app when user uploads the audio then emotion will display on the screen . As we know that this is not a binary classification because of different type of emotions we have decide to use the Multilayer Perceptron Classifier. First we thought to use the SVM(Support Vector Machine) and Multilayer

In this process flow we used neural networks to train the model comes under deep learning side. So we used the MLP Classifier to process the extracted features from the feature extraction process to predict .This model gives us the best results based on the algorithms that we have used in this model and this is basically used for the multi phased models.

The system will predict the emotion based on the audio that we are going to input and this can be implemented in the web application format also. So for this we have used the RAVDEES dataset. This dataset is used by many researches this helped a lot for the researchers who are working on the audio related projects and also initial processing is easy if we use this model Since the model is not a binary classifier because it is not a single plane so we have used the multi classifier known as Multilayer Perceptron so this functionality is basically helps in processing the extracted features from the audio and sends to the model for training.

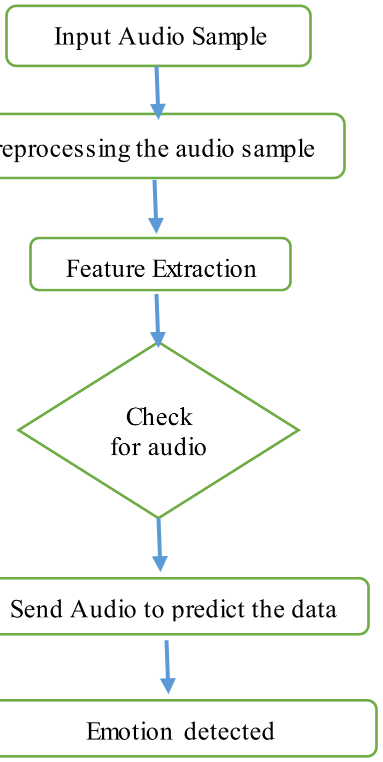


Fig 1: Flow chat of testing process

The literature and related work is represented in the chapter 2 and 3, and the database is defined in the chapter 4,results and outputs are discussed in the . chapter 5, .conclusion and the references are mentioned in below.

II. RELATED WORK

There are many researchers have previously worked on this model by performing different Algorithms with various datasets. Any they have used different classifiers . Authors Welong fu , Wegon performed PNN probabilistic techniques and extracted features from that and that features are passed in to the support vector Machine Classifier (SVM).

In [2] author odoros lliou christos-nikolaos Anagnostopoulos researched and worked three different techniques SVM_(Support vector machine) and MLP_(Multi-Layer Perceptron) and PNN June 10th and finding accuracies between them.

In [3] H.K Palo in their research using the Multilayer Perceptron network in the emotion detection they proposed the method in which features such as MFCC, ,Mel frequency, LPCC,LPC,PLP are extracted from the audio and it is passed in to the Neural Network using MLP_Classifier.

In[9] Yi-Lin and Gang Wei used HMM AND SVM and also in cybernetics in the year of 2005 international conference om machine learning.

In [10] S..Mirsamadi and E. Barsoum ,C. Zhang performed emotion detection using Recurrent Neural Network (RNN) with local attention and it was on the automated speech.

III. PROBLEM STATEMENT

Audio emotions are represented in our audio in the form of tone and pitch. we are going to identify different types of emotions such as sadness ,happiness , neutral ,angry etc.. these emotions are predict using the neural networks so for that after pre-processing the data we are extracted important features from the audio using the feature extraction techniques such as MFCC, Mel frequency and these extracted features are passed into the MLP Classifier for the classification of emotions RAVDEES is the dataset used in this paper and many people have predicted the accuracy by using the different models but in this we thought to build an we application by using the flask framework System model.

A. *Dataset Details*

RAVDESS dataset is a famous dataset which was used by many researchers and it consists of 24 actors and 12 male actors and 12 female actors that are numerically encoded from 01 t0 24 the odd numbered are male ones audio and even indexed audios are female one's the emotions that are present in the dataset are angry, sad, neutral, happy, the data set contains only one format that is audio format fixed audios are pronounced by all the 24 actors for all types of emotion in which each sentence is two times repeated and the intensity of the audio is only in two forms strong and normal it usually consists of 60 trails for all the 24 actors which makes a total of 1440 files . and the data is numerically encoded.

The name of the file is 7_parts the 3rd numerical side of the file represents label to the emotions and those emotions are labelled like '01-neutral' ,'02-happy' etc… The model performance is evaluated using multiple parameters including accuracy, precision, recall, f1-score and support [12-.

B. *MLP and Neural Network Classifier*

This MLP classifier is build up by the perceptron's which contains the input and the output will make the predictions of the passed inputs and middle layers are known as hidden layers where the main functionality goes on there this can be modified as per requirement in this MLP.

In this neural network consists of hundreds of input layers nearly 40k hidden layers and 1 output layer input layer passes the five extracted features from the audio clips the features are extracted from the techniques known as MFCC, Mel frequency ,Chroma, Tonnetz and Cotrast . The activation functions are hidden layers to act up on the input data and to process it .In this we have used the logistic activation function the output of this is the predicted emotion. Figure 1 shows the structure of neural network.

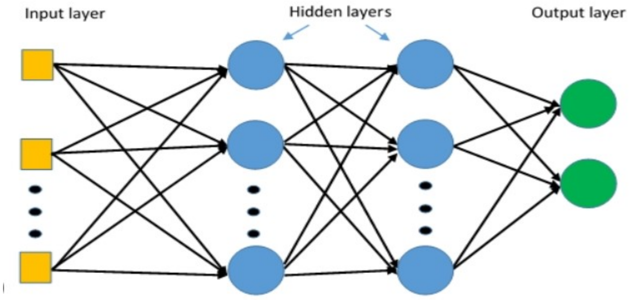


Fig 2: Neural network structure

And this MPL is mainly used for the supervised learning kind of models and it is used for the classification purpose the training invokes the MLP_Classifier to learn the relation between the input and output data in the training process this classifiers adjusts the parameters of the respective model like numerical weights and bias for the purpose of error minimization and the errors can be evaluated in different ways.

This MLP_Classifier implements a MLP_algorithm moreover it uses the backpropagation to train the neural network.

There are some steps that this MLP Classifier involved in Initialization of classifier:

- Initialization of classifier
- Training the given data
- Predicting the output
- Accuracies calculation from the predictions

C. Feature Extraction

Feature extraction is very important for the training because the audio files is processed in the format of extraction and with this feature extraction we will get the main features that are participated in the audio which helps in the prediction of the emotions so there are some techniques to perform this like MFCC , Mel , Chroma, Tonnetz [4].

MFCC

It is one of the feature extraction techniques where it extracts the important features from the audio signals and also it frames the audio signals in the range of 20 to 40 frames suppose if there is a longer frames then signals will change too much throughput [5].

Mel

It logarithmically processes frequencies above a certain threshold (the corner frequency).

Chroma

It is basically a Descriptor_ , which represents the tonal content of a musical audio signal in a condensed mode therefore the features of the Chroma is crucial for semantic analysis like chord recognition.

IV. Implementation

TAfter pre-processing the data and after removing the unwanted and error audio clips from the dataset we appended the correct and clean dataset into the new data frame and then we performed some visualization analysis on the audio files and then we planned to extract the important features from the audio by using feature extraction techniques and then we observed that with only one feature extraction technique we cannot determine the emotion although if determined it will not give that good accuracy so for that we have used the different type of feature extraction techniques the classifier is efficient for the data in our prediction process. Figure 3 shows the workflow of the proposed work.

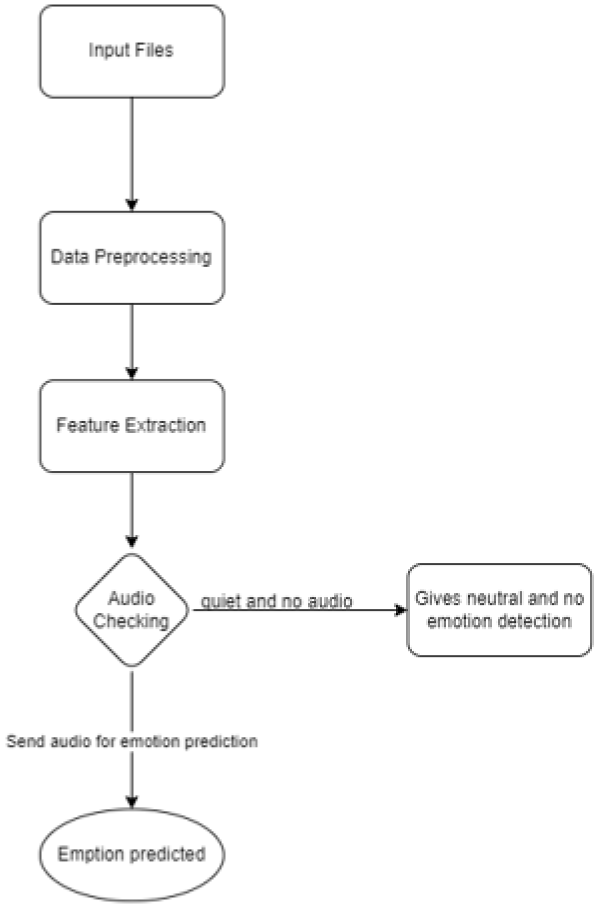


Fig. 3. Workflow of the proposed work

V. Result Analysis

To understand the nature of the audio files we used the spectrograms to find its nature and y – axis represents the Sound Amplitude up to what pitch it is going and X-axis represents the time in seconds. It is one of the virtual way of representing the signal strength and loudness of the audio at different frequencies. Figure 4 and 5 shows the audio Signal Spectrograms.

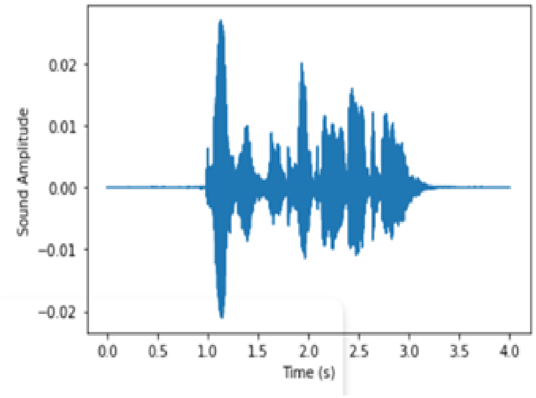


Fig.4. Audio Signal Spectogram

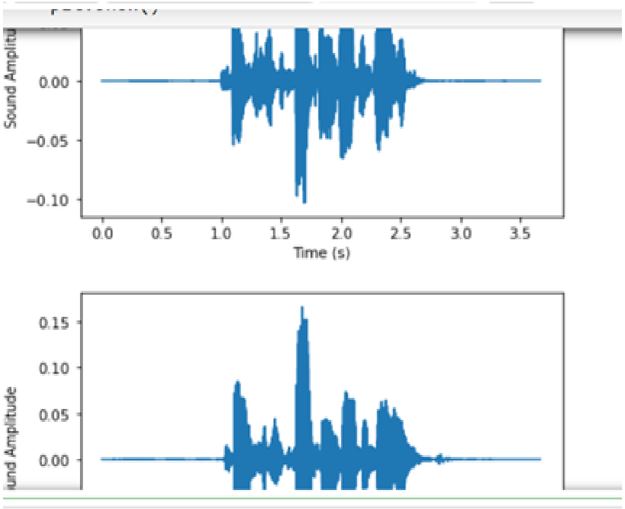


Fig.5. Audio Signal Spectograms

The Above audio spectrograms gives the information about the audio and visual representation of the audio and easy to identify where the pitch is high or low.

A. Amplitude Envelope

This amplitude envelope method gives us a clear idea on at what intensity the waves are going high the red marked line is the pitches up to where the intensity is high so with this we can find at what peek the intensity is more which will be more helpful to our own analysis the main use of this amplitude envelope is it can easily identify the sounds and also uniquely .

Differentiates from other sounds in fig 6 there is a red line in the extremes of the signal it is the envelope in the amplitude. Figure 6 represents audio envelope.

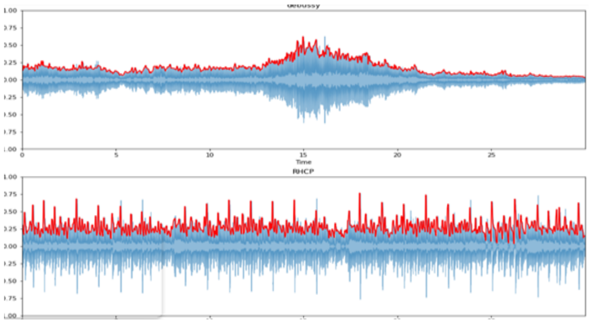
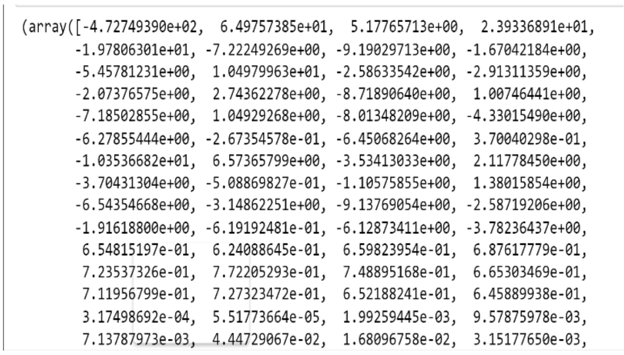


Fig.6. Amplitude Envelope

B. Extracted Features are passed in to the neural network in the form of:

These are in the form of numerical encoding we all know that neural networks will takes only numerical inputs and these are called the feature data. Figure 7 represents extracted features.

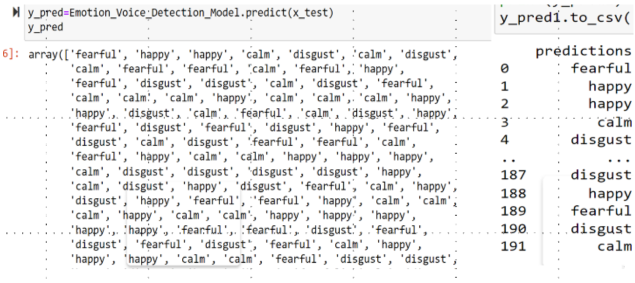


```
(array([-4.72749390e+02,  6.49757385e+01,  5.17765713e+00,  2.39336891e+01,
        -1.97806301e+01, -7.22249269e+00, -9.19029713e+00, -1.67042184e+00,
        -5.45781231e+00,  1.04979963e+01, -2.58633542e+00, -2.91311359e+00,
        -2.07376575e+00,  2.74362278e+00, -8.71890640e+00,  1.00746441e+00,
        -7.18502855e+00,  1.04929268e+00, -8.01348209e+00, -4.33015490e+00,
        -6.27855444e+00, -2.67354578e-01, -6.45068264e+00,  3.70040298e-01,
        -1.03536682e+01,  6.57365799e+00, -3.53413033e+00,  2.11778450e+00,
        -3.70431304e+00, -5.08869827e-01, -1.10575855e+00,  1.38015854e+00,
        -6.54354668e+00, -3.14862251e+00, -9.13769054e+00, -2.58719206e+00,
        -1.91618800e+00, -6.19192481e-01, -6.12873411e+00, -3.78236437e+00,
         6.54815197e-01,  6.24088645e-01,  6.59823954e-01,  6.87617779e-01,
         7.23537326e-01,  7.72205293e-01,  7.48895168e-01,  6.65303469e-01,
         7.11956799e-01,  7.27323472e-01,  6.52188241e-01,  6.45889938e-01,
         3.17498692e-04,  5.51773664e-05,  1.99259445e-03,  9.57875978e-03,
         7.13787973e-03,  4.44729067e-02,  1.68096758e-02,  3.15177650e-03,
```

Fig.7. Extracted Features

C. Predicted outputs

After passing the extracted features in to the neural network model the MLP Classifier will segregate the emotions into their respective categories and then later we got some outputs which shows the emotions of respective audio clips. Predicted outputs are showing in figure 8.



```
y_pred=Emotion_Voice_Detection_Model.predict(x_test)
y_pred

6]: array(['fearful', 'happy', 'happy', 'calm', 'disgust', 'calm', 'disgust',
           'calm', 'fearful', 'fearful', 'calm', 'fearful', 'happy',
           'fearful', 'disgust', 'disgust', 'calm', 'disgust', 'fearful',
           'calm', 'calm', 'calm', 'happy', 'calm', 'calm', 'calm', 'happy',
           'happy', 'disgust', 'calm', 'fearful', 'calm', 'disgust', 'happy',
           'fearful', 'disgust', 'fearful', 'disgust', 'happy', 'fearful',
           'disgust', 'calm', 'disgust', 'fearful', 'fearful', 'calm',
           'fearful', 'happy', 'calm', 'calm', 'happy', 'happy', 'happy',
           'calm', 'disgust', 'disgust', 'disgust', 'disgust', 'happy',
           'calm', 'disgust', 'happy', 'disgust', 'fearful', 'calm', 'happy',
           'disgust', 'happy', 'fearful', 'fearful', 'happy', 'calm', 'calm',
           'calm', 'happy', 'calm', 'calm', 'happy', 'happy', 'happy',
           'happy', 'happy', 'fearful', 'fearful', 'disgust', 'fearful',
           'disgust', 'fearful', 'disgust', 'fearful', 'calm', 'happy',
           'happy', 'happy', 'calm', 'calm', 'fearful', 'disgust', 'disgust',
```

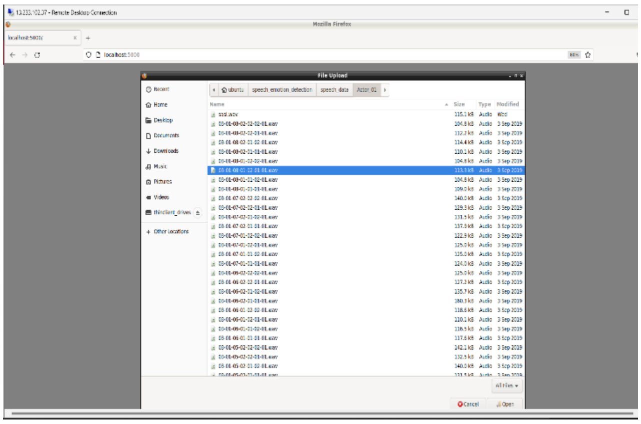
```
y_pred1.to_csv(

     predictions
0        fearful
1          happy
2          happy
3           calm
4        disgust
..           ...
187      disgust
188        happy
189      fearful
190      disgust
191         calm
```

Fig.8. Predicted Outputs

D. Accuracy calculation

After predicting the output we tried to find accuracy and we got accuracy up to 70 percent and then after some discussions we planned to discuss some outcomes in the format of web application. So that it will become more meaningful and we performed this by using the flask app and this can be done in Ubuntu which is known as operating system and we kept the database in local system. Figure 9 shows audio selection.



Name	Size	Type	Modified
sad.wav	115.1 kB	Audio	Wed
03-01-08-02-02-02-01.wav	104.8 kB	Audio	3 Sep 2019
03-01-08-02-02-01-01.wav	112.2 kB	Audio	3 Sep 2019
03-01-08-02-01-02-01.wav	114.4 kB	Audio	3 Sep 2019
03-01-08-02-01-01-01.wav	110.1 kB	Audio	3 Sep 2019
03-01-08-01-02-02-01.wav	104.8 kB	Audio	3 Sep 2019
03-01-08-01-02-01-01.wav	113.3 kB	Audio	3 Sep 2019
03-01-08-01-01-02-01.wav	104.8 kB	Audio	3 Sep 2019
03-01-08-01-01-01-01.wav	109.0 kB	Audio	3 Sep 2019
03-01-07-02-02-02-01.wav	140.0 kB	Audio	3 Sep 2019
03-01-07-02-02-01-01.wav	129.3 kB	Audio	3 Sep 2019
03-01-07-02-01-02-01.wav	131.5 kB	Audio	3 Sep 2019
03-01-07-02-01-01-01.wav	137.9 kB	Audio	3 Sep 2019
03-01-07-01-02-02-01.wav	122.9 kB	Audio	3 Sep 2019
03-01-07-01-02-01-01.wav	125.0 kB	Audio	3 Sep 2019
03-01-07-01-01-02-01.wav	125.0 kB	Audio	3 Sep 2019
03-01-07-01-01-01-01.wav	124.0 kB	Audio	3 Sep 2019
03-01-06-02-02-02-01.wav	125.0 kB	Audio	3 Sep 2019
03-01-06-02-02-01-01.wav	127.2 kB	Audio	3 Sep 2019
03-01-06-02-01-02-01.wav	135.7 kB	Audio	3 Sep 2019
03-01-06-02-01-01-01.wav	180.1 kB	Audio	3 Sep 2019
03-01-06-01-02-02-01.wav	118.6 kB	Audio	3 Sep 2019
03-01-06-01-02-01-01.wav	110.1 kB	Audio	3 Sep 2019
03-01-06-01-01-02-01.wav	116.5 kB	Audio	3 Sep 2019
03-01-06-01-01-01-01.wav	117.6 kB	Audio	3 Sep 2019
03-01-05-02-02-02-01.wav	142.1 kB	Audio	3 Sep 2019
03-01-05-02-02-01-01.wav	132.5 kB	Audio	3 Sep 2019
03-01-05-02-01-02-01.wav	140.0 kB	Audio	3 Sep 2019
03-01-05-02-01-01-01.wav	131.5 kB	Audio	3 Sep 2019

Fig.9. Selecting the audio

So basically when user tries to upload an audio clip from the local system after uploading the file then the training of the model starts and it will predict the emotion in the text format that would be displayed in the web application.

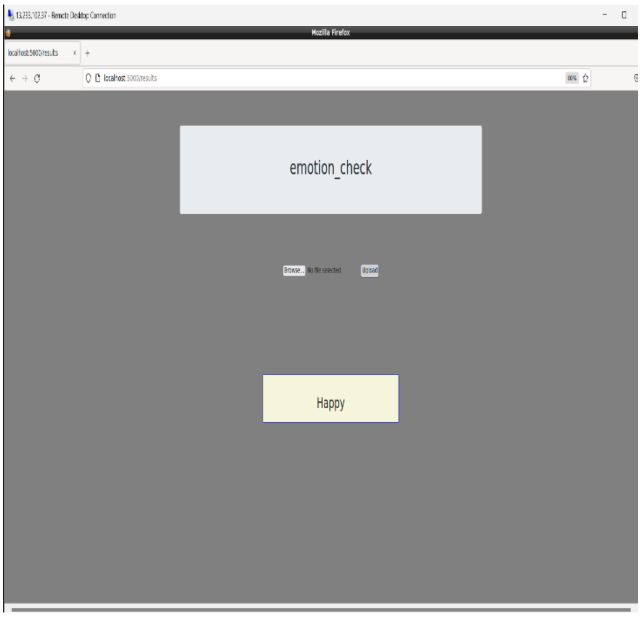
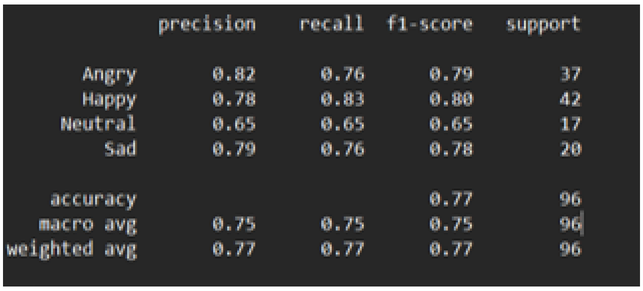


Fig. 10. Mobile application

And this web application is basically designed by using the flask framework and we used the Ubuntu operating system and this is speech emotion detection model and web application it will trains the MLP model which consists of RAVDEES dataset and this model helps us in predicting the emotion and with help of html and css we designed the application and we used some in build libraries to build this application they are: Librosa, sklearn, pyaudio, numpy for installation of flask we used Flak library.

E. Accuracy comparision

Figure 11 representing performance report of MLP classifier on speech emotion detection based on audio processing. This model is giving 79% accuracy and training accuracy is 82% and testing accuracy is 76%. The figure showing the other parameters like precision, recall, f1-score and support parameters values of MLP classifier.



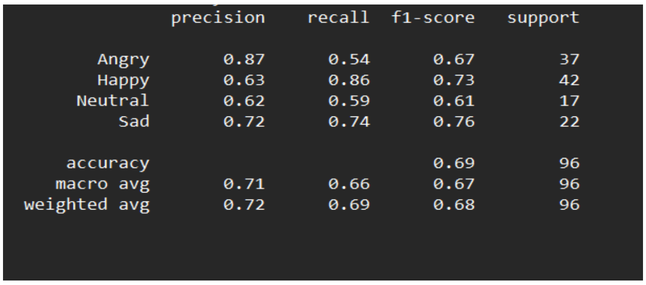
```
              precision    recall  f1-score   support

       Angry       0.82      0.76      0.79        37
       Happy       0.78      0.83      0.80        42
     Neutral       0.65      0.65      0.65        17
         Sad       0.79      0.76      0.78        20

    accuracy                           0.77        96
   macro avg       0.75      0.75      0.75        96
weighted avg       0.77      0.77      0.77        96
```

Fig. 11. Classification Report of MLP_Classifier

Figure 12 representing performance report of SVM classifier on speech emotion detection based on audio processing. This model is giving 72% accuracy and training accuracy is 78.60% and testing accuracy 68.75% is The figure showing the other parameters like precision, recall, f1-score and support parameters values of SVM classifier.



```
              precision    recall  f1-score   support

       Angry       0.87      0.54      0.67        37
       Happy       0.63      0.86      0.73        42
     Neutral       0.62      0.59      0.61        17
         Sad       0.72      0.74      0.76        22

    accuracy                           0.69        96
   macro avg       0.71      0.66      0.67        96
weighted avg       0.72      0.69      0.68        96
```

Fig. 12. Classification Report of SVM_Classifier

The performance of MLP and SVM classifiers are compared in terms of model performance metrics. The analysis on accuracy parameter is shown in table 1.

TABLE I. PERFORMANCE ANALYSIS

Classifier	***Train accuracy***	***Test accuracy***	***Model accuracy***
MLP	82%	76%	79%
SVM	78.60%	68.75%	72%

VI. CONCLUSION

Finally we concluded with the above results that predicting the emotion and finding the accuracy of the predicted results and what are feature extraction techniques that are participated in the process of prediction and we came to know that what are the different Classfiers that we can use in the prediction of emotion and also which Classfier is best to perform this and in this paper we used the MLP Classifier to process the extracted features .So finally there is a further scope to this project like there is chance of implementing it more dynamically.

VII. FUTURE WORK

As part of future work, the performance of the proposed work is going to compare with other achine learning models and neural network models. Finally, we are planning to introduce an ensemble model to detect emotion through audio processing.

REFERENCES

[1] Hassan, M. M., Alam, M. G. R., Uddin, M. Z., Huda, S., Almogren, A., & Fortino, G. (2019). Human emotion recognition using deep belief network architecture. Information Fusion, 51, 10-18.

[2] Iliou, T., & Anagnostopoulos, C. N. (2010, June). SVM-MLP-PNN classifiers on speech emotion recognition field-A comparative study. In 2010 Fifth International Conference on Digital Telecommunications (pp. 1-6). IEEE.

[3] Palo, H. K., Mohanty, M. N., & Chandra, M. (2015). Use of different features for emotion recognition using MLP network. In Computational Vision and Robotics (pp. 7-15). Springer, New Delhi.

[4] Kattel, M., Nepal, A., Shah, A. K., & Shrestha, D. (2019, January). Chroma feature extraction. In Conference: chroma feature extraction using fourier transform (No. 20).

[5] Alim, S. A., & Rashid, N. K. A. (2018). Some commonly used speech feature extraction algorithms (pp. 2-19). London, UK:: IntechOpen.

[6] Davis, S.Mermelstein, P.(1980) Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. In IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28 No. 4, pp. 357- 366.

[7] Chen, Jianxin, et al. "Predicting syndrome by NEI specifications: a comparison of five data mining algorithms in coronary heart disease." International Conference on Life System Modeling and Simulation. Springer, Berlin, Heidelberg, 2007.

[8] X. Huang, A. Acero, and H. Hon. Spoken Language Processing: A guide to theory, algorithm, and system development. Prentice Hall, 2001

[9] Yi-Lin Lin and Gang Wei, Speech_emotion_recognition based on HMM and SVM 2005 International Conference on Machine Learning and Cybernetics, 2005, pp. 4898-4901 Vol. 8, doi: 10.1109/ICMLC.2005.1527805.

[10] S. Mirsamadi, E. Barsoum and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 2227- 2231, doi: 10.1109/ICASSP.2017.7952552.

[11] Mustaqeem, M. Sajjad and S. Kwon.Speech Emotion Recognition based on Clustering by analyzing the Learned Features and Deep BiLSTM in IEEE Access, vol. 8, pp. 79861-79875, 2020, doi: 10.1109/ACCESS.2020.2990405.

[12] Kavitha, M., Srinivasulu, S., Madhava Reddy, M., Gopikrishna, V., Phani Kumar, S., & Kavitha, S. (2022). Hybrid Model Using Feature Selection and Classifier in Big data Healthcare Analytics. In Inventive Communication and Computational Technologies (pp. 777-791). Springer, Singapore.

[13] Kavitha, M., Srinivas, P. V. V. S., Kalyampudi, P. L., & Srinivasulu, S. (2021, September). Machine Learning Techniques for Anomaly Detection in Smart Healthcare. In 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 1350-1356). IEEE.