National University of Computer and Emerging Sciences FAST

Data Science Project

Spring 2024

Project Report: **Speech Emotion Recognition**

BSCS-6B

Instructors: **Dr. Muhammad Nouman Durrani**

Team Members:

21K4556 Muhammad Anas

21K3323 Muhammad Shaheer

Introduction:

In this project, we aim to develop a Speech Emotion Recognition (SER) system using the Toronto Emotional Speech Set (TESS) dataset. The TESS dataset comprises 2800 audio recordings of individuals speaking various sentences with six different emotions: anger, disgust, fear, happiness, neutral, and sadness. The primary goal is to build a machine learning model capable of accurately predicting the emotional state of an individual based on their speech.

Data Preprocessing:

To prepare the data for training, the following steps were performed:

Data Loading: The TESS dataset was loaded using the Librosa library, which is widely used for audio analysis tasks.

Feature Extraction: Various features were extracted from each audio sample, including:

Mel-frequency cepstral coefficients (MFCCs)

Chroma features

Spectral contrast These features are essential for capturing different aspects of the audio signal and are commonly used in speech recognition tasks.

Dataset: Toronto Emotional Speech Set (TESS)

Link: TESS Dataset on Kaggle

Model Architecture:

The proposed model architecture consists of a combination of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) layers:

CNN Layers: CNN layers are employed to extract spatial features from the input feature matrix (MFCCs, chroma features, etc.).

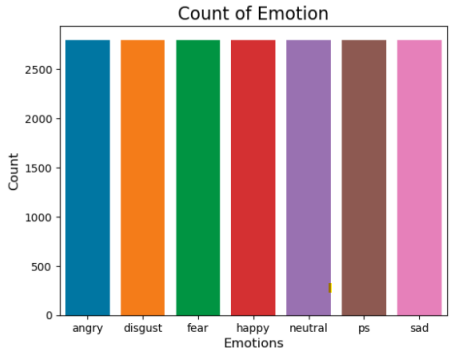
LSTM Layers: LSTM layers are utilized to model temporal dependencies and capture sequential patterns in the data. They are particularly effective for processing sequential data, such as speech signals, due to their ability to capture long-range dependencies and handle variable-length sequences.

Fully Connected Layer: Following the CNN-LSTM layers, a fully connected layer is added for classification. This layer performs emotion classification based on the extracted features, with a SoftMax activation function applied to obtain a probability distribution over the different emotion classes.

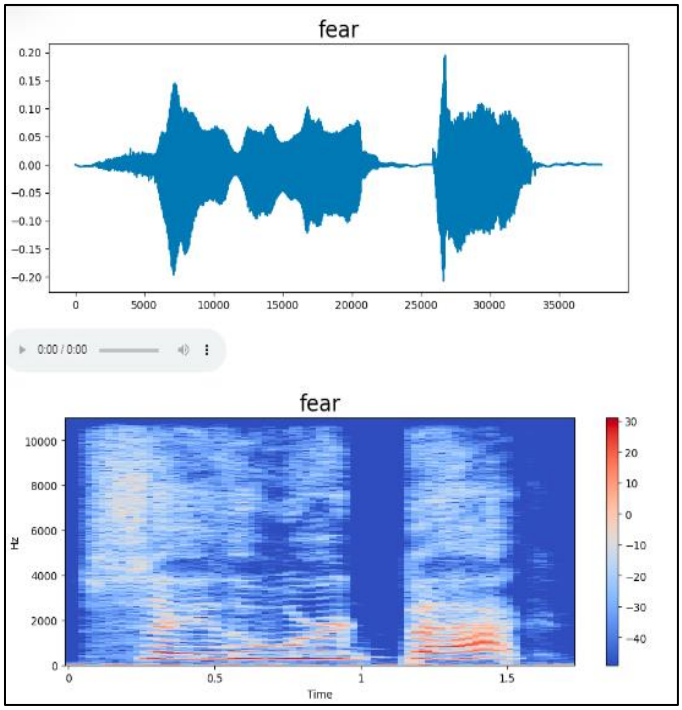
Training and Evaluation:

The TESS dataset was split into training, validation, and test sets, with 80%, 10%, and 10% of the data allocated to each set, respectively. The model was trained on the training set for 50 epochs while monitoring accuracy and loss on the validation set to prevent overfitting. Once the model converged, its performance was evaluated on the test set.

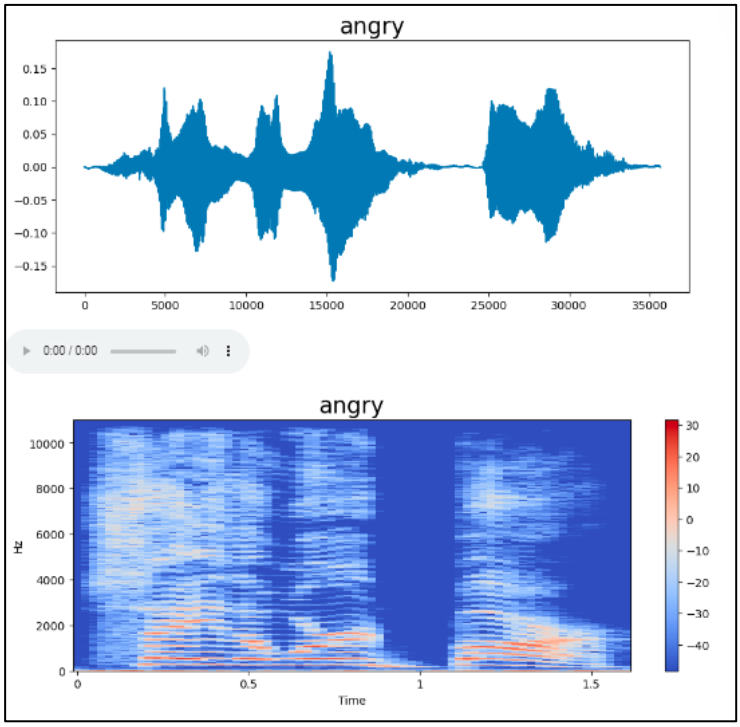
Results:



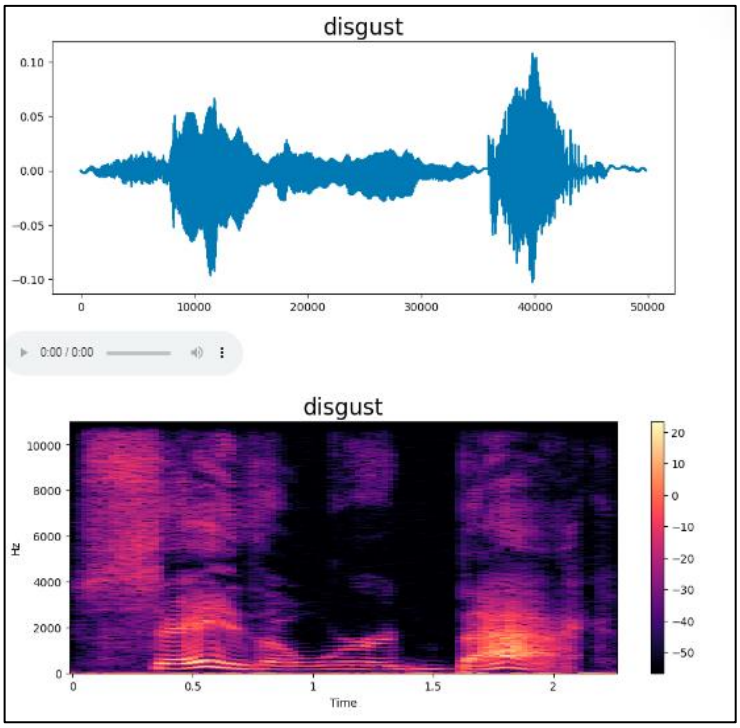
Fear:



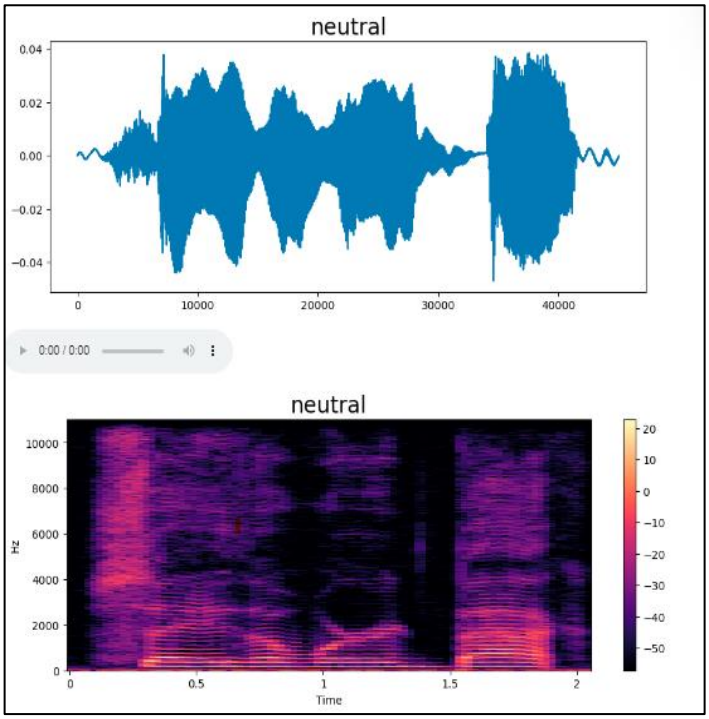
Angry:



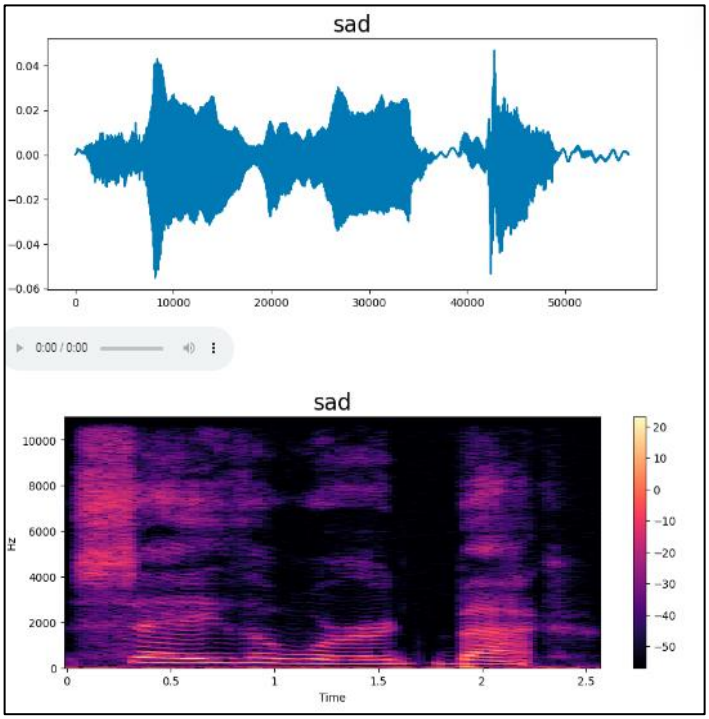
Disgust:



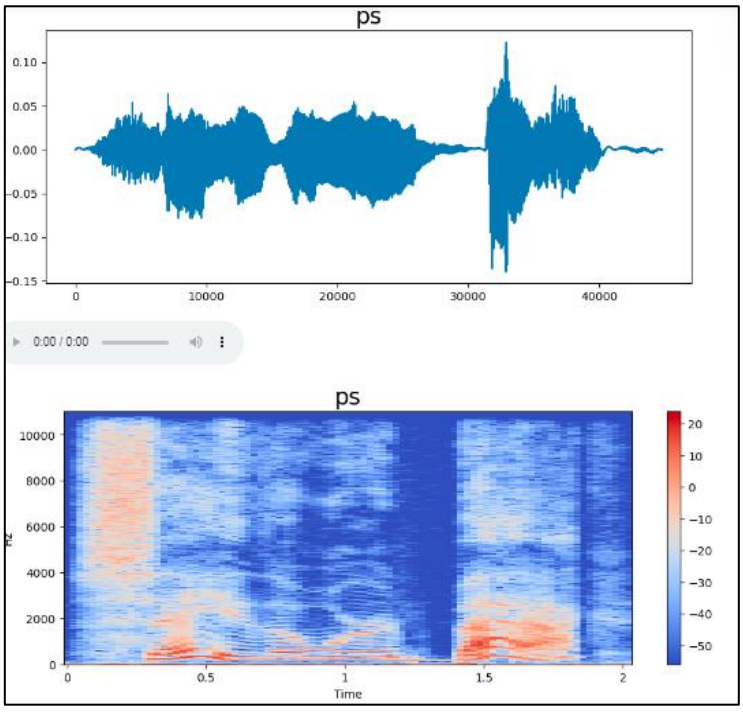
Neutral:


neutral
0.04
0.02
0.00
−0.02
−0.04
0
10000
20000
30000
40000
0:00 / 0:00
neutral
10000
8000
6000
4000
2000
0
Hz
0
0.5
1
1.5
2
Time
20
10
0
−10
−20
−30
−40
−50

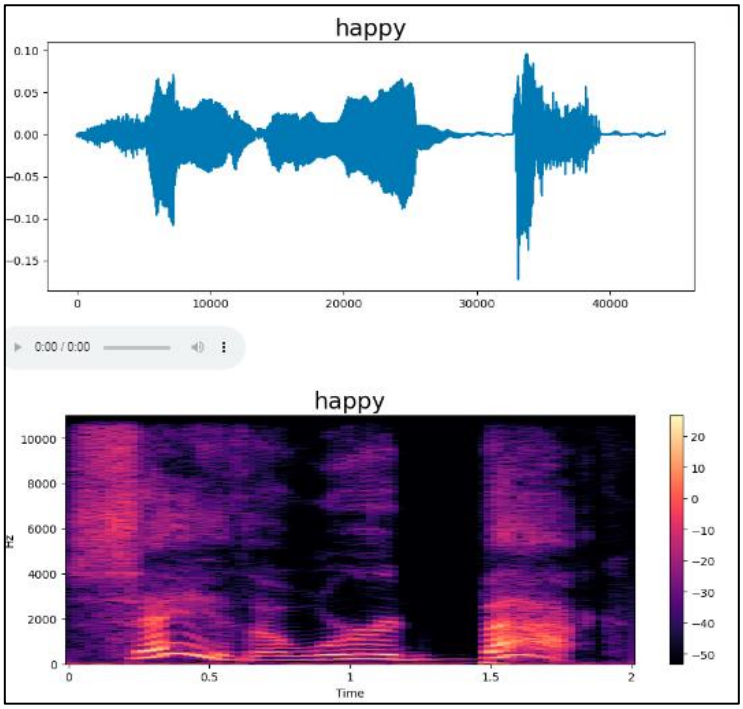
Sad:


sad
0.04
0.02
0.00
−0.02
−0.04
−0.06
0
10000
20000
30000
40000
50000
0:00 / 0:00
sad
10000
8000
6000
4000
2000
0
Hz
0
0.5
1
1.5
2
2.5
Time
20
10
0
−10
−20
−30
−40
−50

Ps:


ps
0.10
0.05
0.00
−0.05
−0.10
−0.15
0
10000
20000
30000
40000
0:00 / 0:00
ps
10000
8000
6000
4000
2000
0
Hz
0
0.5
1
1.5
2
Time
20
10
0
−10
−20
−30
−40
−50

Happy:


happy
0.10
0.05
0.00
−0.05
−0.10
−0.15
0
10000
20000
30000
40000
0:00 / 0:00
happy
10000
8000
6000
4000
2000
0
Hz
0
0.5
1
1.5
2
Time
20
10
0
−10
−20
−30
−40
−50

Training:

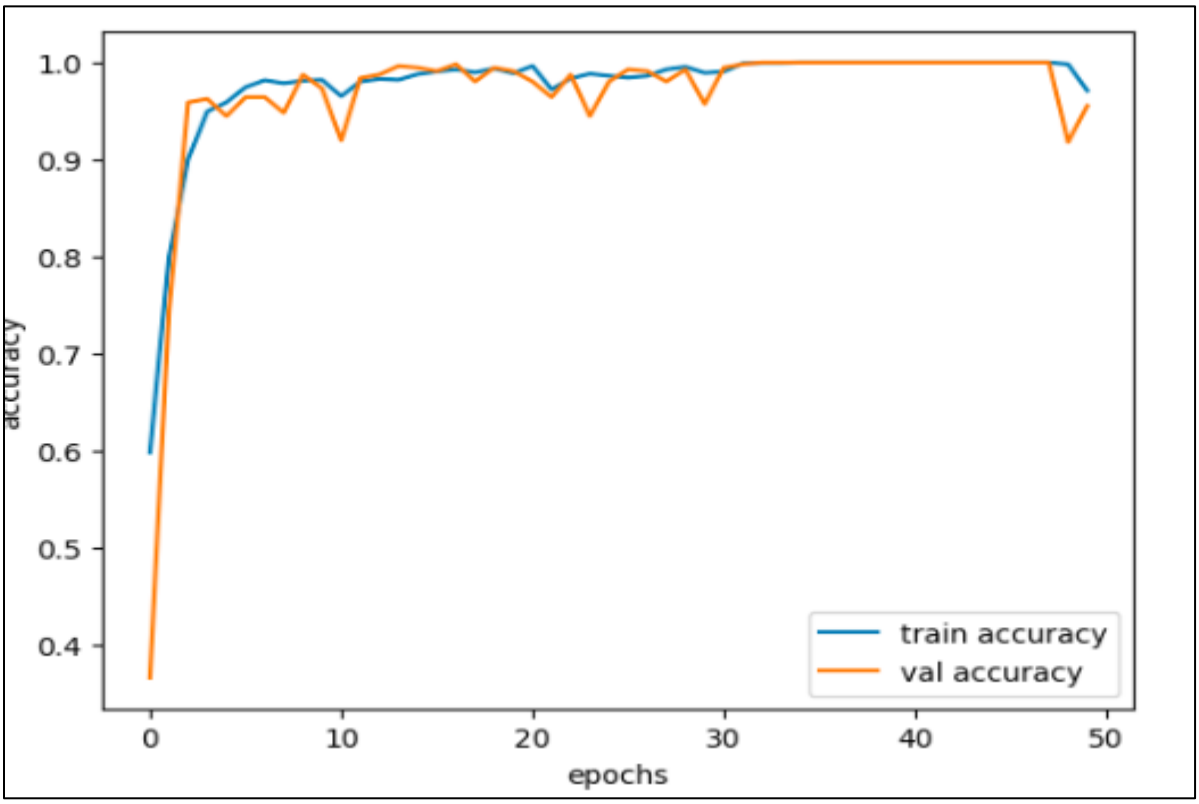
Model: "sequential"

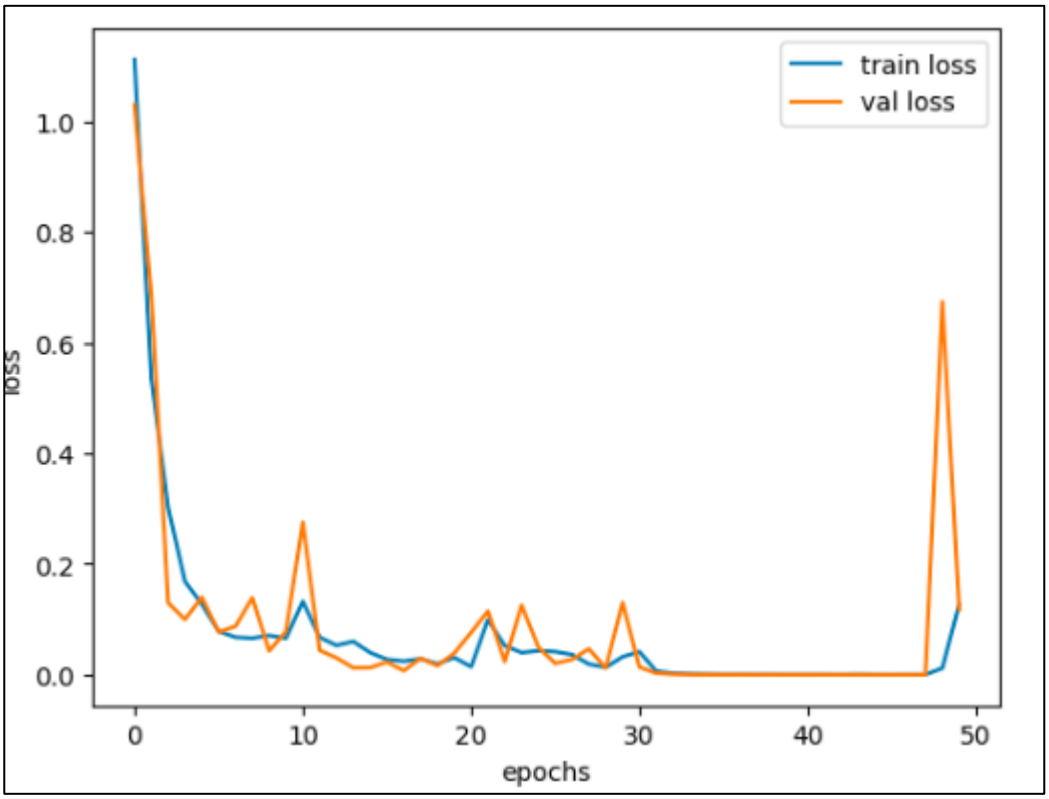
Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 256)	264,192
dropout (Dropout)	(None, 256)	0
dense (Dense)	(None, 128)	32,896
dropout_1 (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 64)	8,256
dropout_2 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 7)	455

Total params: 305,799 (1.17 MB)

Trainable params: 305,799 (1.17 MB)

Non-trainable params: 0 (0.00 B)



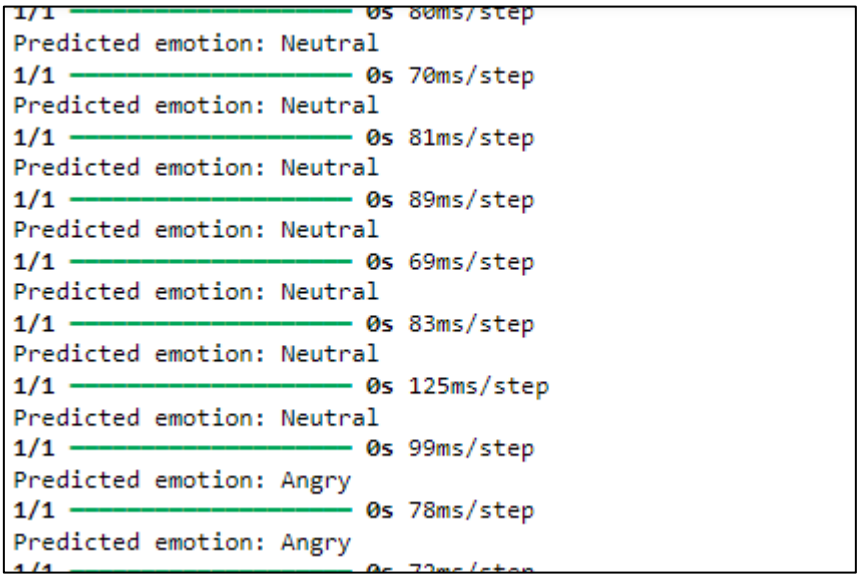


Prediction:

Pre-Trained Data:

```
Predicted emotion for OAF_learn_angry.wav: Neutral
Predicted emotion for OAF_lease_angry.wav: Neutral
Predicted emotion for OAF_lid_angry.wav: Disgust
Predicted emotion for OAF_life_angry.wav: Disgust
Predicted emotion for OAF_limb_angry.wav: Disgust
Predicted emotion for OAF_live_angry.wav: Neutral
Predicted emotion for OAF_loaf_angry.wav: Disgust
Predicted emotion for OAF_long_angry.wav: Neutral
Predicted emotion for OAF_lore_angry.wav: Neutral
Predicted emotion for OAF_lose_angry.wav: Neutral
Predicted emotion for OAF_lot_angry.wav: Disgust
Predicted emotion for OAF_love_angry.wav: Disgust
Predicted emotion for OAF_luck_angry.wav: Disgust
Predicted emotion for OAF_make_angry.wav: Neutral
Predicted emotion for OAF_match_angry.wav: Disgust
Predicted emotion for OAF_merge_angry.wav: Neutral
Predicted emotion for OAF_mess_angry.wav: Neutral
Predicted emotion for OAF_met_angry.wav: Disgust
Predicted emotion for OAF_mill_angry.wav: Neutral
```

Trained Data Output:



Usages:

1. Security and Surveillance
2. Entertainment
3. Education
4. HealthCare
5. Call Centers and Customer Service Chatbots
6. Mental Health Monitoring
7. Speech Therapy

Libraries Used:

Os Keras Numpy Matplotlib

Librosa Tensorflow Pyaudio

Sklearn Pandas Seaborn Ipython Display

Conclusion:

In summary, we have developed a Speech Emotion Recognition system using the TESS dataset and a CNN-LSTM model. The model demonstrated high accuracy on the test set, indicating its ability to accurately predict the emotional state of individuals based on their speech. This system holds promise for a wide range of applications, including customer service chatbots, mental health monitoring, and speech therapy.