A

PROJECT REPORT

ON

A CLINICAL APPLICATION FOR SPEECH EMOTION RECOGNITION

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE OF

BACHELOR OF ENGINEERING
INFORMATION TECHNOLOGY

BY

Pushkar Kane	B190058593
Pratik Mathe	B190058633
Yash Waghumbare	B190058743

Under the guidance of
Mr. Ganesh Pise

Department Of Information Technology
Pune Institute of Computer Technology
Pune - 411 043.
2022-2023

SCTR's PUNE INSTITUTE OF COMPUTER TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY

CERTIFICATE

This is to certify that the final project report entitled
A Clinical Application for Speech Emotion Recognition
submitted by

Pushkar Kane	B190058593
Pratik Mathe	B190058633
Yash Waghumbare	B190058743

is a bonafide work carried out by them under the supervision of **Mr. Ganesh Pise** and it is approved for the partial fulfillment of the requirement of Savitribai Phule Pune University for the award of the Degree of Bachelor of Engineering (Information Technology).

This project report has not been earlier submitted to any other Institute or University for the award of any degree or diploma.

Mr. Ganesh Pise
Project Guide

Dr. A. S. Ghotkar
HOD IT

SPPU External Guide

Dr. S. T. Gandhe
Principal

Date:
Place:

Acknowledgement

Purpose of acknowledgments page is to show appreciation to those who contributed in conducting this dissertation work / other tasks and duties related to the report writing. Therefore when writing acknowledgments page you should carefully consider everyone who helped during research process and show appreciation in the order of relevance. In this regard it is suitable to show appreciation in brief manner instead of using strong emotional phrases.

In this part of your work it is normal to use personal pronouns like “I, my, me” while in the rest of the report this articulation is not recommended. Even when acknowledging family members and friends make sure of using the wording of a relatively formal register.

The list of the persons you should acknowledged, includes guide (main and second), head of dept, academic staff in your department, technical staff, reviewers, head of institute, companies, family and friends.

You should acknowledge all sources of funding. It’s usually specific naming the person and the type of help you received. For example, an advisor who helped you conceptualize the seminar,someone who helped with the actual building or procedures used to complete the seminar,someone who helped with computer knowledge, someone who provided raw materials for the seminar, etc.

Pushkar Kane	B190058593
Pratik Mathe	B190058633
Yash Waghumbare	B190058743

Abstract

Presently AI, is used in various medical fields. Mental health in an important part of overall health of a person and speech is the primary form for expression of emotion. Thus, Speech Emotion Recognition can be used to understand emotions of the person and help doctors focus on the cure. Speech Emotion recognition is analysis and classification of speech signals to detect the underlying emotions. This paper proposes a model for speech emotion recognition using an attention mechanism. The model is developed using Mel-frequency cepstral coefficients (MFCCs), and a combination of 2D CNN layers and LSTM recurrent layers for temporal aggregation. The proposed model is evaluated using a dataset of speech recordings containing eight emotion categories. The results show that the model achieves 89.93% accuracy. The attention mechanism is found to improve the recognition performance by focusing on relevant emotional information and ignoring irrelevant information. This research has potential applications in clinical settings in detection as well as treatment for mental health issues.

Keywords: Speech Emotion Recognition (SER), Mental Health, Deep Learning, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Mel Frequency Cepstrum Coefficients (MFCC), Attention Mechanism, RAVDESS,

Contents

List of Figures

Abbreviations

SER	:	Speech Emotion Recognition
CNN	:	Convolutional Neural Network
MFCC	:	Mel frequency cepstral coeffi- cient
DNN	:	Deep Neural Network
LSTM	:	Long short-term memory

1. Introduction

1.1 Introduction

Monitoring and treating mental health are crucial for overall physical health and a safer community and social life [1], as overall globally over 1.1 billion individuals were diagnosed with mental disorders in 2016 [2]. The COVID-19 pandemic has exacerbated the situation, with depression and anxiety disorders increasing by 25% globally during the first year, particularly among young people and women [3]. Unfortunately, late or unreceived mental care has led to an increase in suicides, with one person dying by suicidal action related to a mental disorder every 40 seconds [4]. Furthermore, the United States has experienced over 200 cases of mass shootings in less than the first half of the year [5].

In speech, individuals express common emotions such as happiness, sadness, anger, worry, fear, and neutrality. Changes in an individual's emotions can be indicative of certain mental disorders such as depression, mood disorders and trauma and stress disorders. Early diagnosis of mental disorders is crucial to provide the correct treatment and preventing severe illnesses and suicidal actions. However, the current screenings for these diseases rely mainly on psychological examination and interviews, which can lack objectivity. Therefore, there is a need for technology that can detect psychiatric changes in patients earlier.

A therapist needs to understand the patient's mental state by recognizing emotions through their vocal responses. Based on this, therapists then need to decide the course of treatment by analyzing the patient's progress by comparing his responses in earlier therapeutic settings. Here it becomes challenging for the therapist to keep track of the progress of multiple patients and quantify the improvements in their responses. Thus, a platform can be useful to provide insights and keep track of patient emotional progress by recognizing the embedded emotions in patients' speech in various medical fields and therapy sessions.

AI can relieve doctors of the burden of comprehending their patients' emotional states and move the emphasis from transactional chores to individualized medical treatment and service. However, it necessitates that computers cleverly deduce human speech and comprehend it on a semantic level. Currently, artificial intelligence (AI) is empowering a variety of medical applications, including early detection of diseases, drug discovery, heart disease prediction, and Robot-assisted surgery. Understanding emotions can improve human-machine interaction by enabling machines to understand human behavior better and respond accordingly. Thus, various efforts are being taken in this domain.

SER is the quickest means of communication and information exchange between people and computers, and it has various practical applications in human-computer interaction. Systems for detecting the embedded emotions in voice signals are referred to as Speech Emotion Recognition (SER) systems. Speech Emotion Recognition uses extracted features from raw audio waves using various techniques like MFCC and Log-Mel-Spectrogram. These features can be temporal or spectral features.

In this paper, we propose the use of MFCC technique for feature extraction. The primary goal of SER systems is to identify certain speaker voice traits under various emotional states. Speech emotion recognition technology can be used to detect, monitor and treat mental disorders such as depression, trauma, stress and bipolar disorder. By analyzing a patient's speech patterns, the technology can identify signs of emotional distress, which can be used to inform treatment and intervention plans.

Speech emotion recognition technology can also be used to diagnose autism spectrum disorder (ASD). Individuals affected with autism often have difficulty recognizing and interpreting emotions in speech, and this technology can help clinicians assess these abilities in a standardized and objective way. Parkinson's disease can affect speech patterns, leading to changes in pitch, volume, and articulation. Speech emotion recognition technology can be applied to detect these changes, helping in early detection and treatment of Parkinson's disease.

Individuals who have suffered a stroke may experience speech difficulties, including problems with emotional expression. Speech emotion recognition technology can also be employed in this case for monitoring progress during stroke rehabilitation and providing targeted interventions to improve emotional expression in speech. Overall, speech emotion recognition technology has the potential to revolutionize clinical diagnosis and treatment by providing objective, standardized assessments of emotional expression in speech.

The objective of this research is to develop a model for accurately detecting and classifying emotions from speech signals using deep learning techniques incorporating attention mechanisms in the model to improve recognition performance by focusing on relevant emotional information and ignoring irrelevant information. Further, we evaluate the built model and compare it with existing model's performances and explore the potential applications of the proposed model in clinical settings for the detection and treatment of mood disorders. We also focus on contributing to the development of AI-based tools for personalized care and medical services, thereby freeing up physicians' tasks of understanding the emotional space of their patients.

1.2 Motivation

The goal of speech emotion recognition (SER) is to identify emotion in speech, regardless of the semantic content. However, since emotions are arbitrary, it might be challenging to record them in everyday speech, even for humans.

However, due to the complexity of emotions, which makes them challenging to perceive, the existing SER models still do not reliably understand human emotions.

Despite the fact that this profession has been present for a while, new developments have brought it back into the spotlight. focused on use in medical care, which can be utilized to better medical care by relating the interaction between sickness and interaction between doctors and patients at the time of a visit.

1.3 Objectives

The goal of speech emotion recognition (SER) is to identify emotion in speech, regardless of the semantic content. However, since emotions are arbitrary, it might be challenging to record them in everyday speech, even for humans. However, due to the complexity of emotions, which makes them challenging to perceive, the existing SER models still do not reliably understand human emotions. Despite the fact that this profession has been present for a while, new developments have brought it back into the spotlight. focused on use in medical care, which can be utilized to better medical care by relating the interaction between sickness and interaction between doctors and patients at the time of a visit.

1.4 Scope

Speech emotion recognition will soon be included into clinical perception in order to assist physicians in monitoring patients' conditions while they are receiving therapy.

The project's scope is as follows:

- Real-time software.
- easy-to-use application
- removes human involvement when assessing the patient's mental state.

2. Literature Survey

2.1 Existing Methodologies

One of the main challenges in SER is the variability of emotions conveyed through speech, which can be influenced by various factors such as cultural background, gender, age, and speaking style. In addition, there is often a lack of labeled speech data for emotion recognition, which makes it difficult to train accurate machine learning models. There are several approaches to feature extraction for SER, including spectral, prosodic, and lexical features. Spectral features involve analyzing the frequency spectrum of speech signals, while prosodic features involve analyzing the rhythm, intonation, and stress patterns of speech. Lexical features involve analyzing the content and language of speech.

Prior to the advent of deep learning, speech emotion recognition (SER) research relied heavily on use of manually engineered audio features and old machine learning models such as Hidden Markov Models (HMMs) and support vector machines (SVMs) [6]. K. Han along with his colleagues introduced the first deep learning-based model [7] for SER in 2014. Their approach utilized deep neural networks to automatically extract complex features from the data, demonstrating the effectiveness of this method for SER. Specifically, they utilized Mel-frequency cepstral coefficients (MFCCs) to extract features and classified audio files into five different emotions. In [8] various statistical features of pitch, energy, and zero crossing rate (ZCR) as well as Mel frequency cepstral coefficients (MFCC) from a database of 2000 utterances are explored. The pitch feature was extracted using AMDF and employed a Naive Bayes classifier for classifying audio signal.

In [9] the author utilized recurrent neural networks (RNNs) on the IEMOCAP dataset and employed raw and emotional low-level descriptors (LLDs) for feature extraction. The resulting accuracy rate was found to be 66.23%, on average. Fatemeh Noroozi's research in 2017[10] presented a vocal-based approach that utilized decision trees and random forest algorithms. The average accuracy rate achieved by this method was 66.28%.

In 2017, a study on speech emotion recognition using spectrograms with deep convolutional neural networks (CNNs) was conducted. The proposed CNN model consisted of three convolutional layers and three fully connected layers.[11]. The authors of [12] have investigated a model which integrates both temporal and spatial features. They expanded the RAVDESS dataset four-fold by adding Additive Gaussian White Noise (AWGN) to 5760 audio samples. To classify emotions from one of the eight classes, the authors created two parallel CNNs for extracting both spatial as well as temporal elements.

MFCC is commonly employed for analyzing audio signals and have shown superior performance for speech-centered emotion recognition methods in comparison to other techniques. In 2020[13], Mustaqeem introduced a CNN model with a deep bidirectional long short-term memory (LSTM) that incorporated MFCC and outperformed the existing models on the IEMOCAP dataset. To ascertain the performance of the advanced DSCNN and CNN models[14], the author conducted two experiments. The experiments resulted in an accuracy rate of 79.4%.

Speech emotion recognition technology can be used to diagnose mental issues of depression, stress, and bipolar disorder. In [15] authors have concluded that by analyzing a patient's speech patterns, the technology can identify signs of emotional distress, which can be used to inform treatment and intervention plans. Parkinson's disease can affect speech patterns, leading to changes in pitch, volume, and articulation. In [16] Tsanas have described that SER technology can be employed to detect these changes, providing early detection and treatment of Parkinson's disease. Authors in [17] demonstrate that individuals who have suffered a stroke may experience speech difficulties, including problems with emotional expression. SER can also be applied in monitoring progress during stroke rehabilitation and provide targeted interventions to improve emotional expression in speech. PTSD can affect speech patterns, leading to changes in tone, pitch, and vol¬¬ume. In [18] "Using interpretable machine learning models to improve PTSD diagnosis" authors suggest that SER technology can be utilized in monitoring improvement during PTSD treatment and provide targeted interventions to improve emotional expression in speech.

Overall, the field of SER is still evolving, and there is much room for more research and experimentation for improving accuracy and making emotion recognition systems more reliable. However, the potential applications of SER in various fields make it an area of research that is likely to continue to receive significant attention in the future. Speech emotion recognition technology has the potential to revolutionize clinical diagnosis and treatment by providing objective, standardized assessments of emotional expression in speech

2.2 Research Gap Analysis

The encoder-decoder paradigm for machine translation was enhanced by the addition of the attention mechanism. The purpose of the attention mechanism was to allow the decoder to use the most pertinent portions of the input sequence in a flexible way by combining all of the encoded input vectors in a weighted manner, with the most pertinent vectors receiving the highest weights. Since then, various machine learning tasks, including key-term extraction, image classification, etc., have used attention methods. The attention mechanism can be employed in the speech emotion-recognition task to direct the model's attention to the parts that can convey emotional information more effectively, ignore irrelevant information, and enhance recognition performance.

3. Requirement Specification and Analysis

3.1 Problem Definition

The Problem is to build a clinical application using Deep learning techniques that can detect patient's emotions embedded in speech and help better interpretation of patient's feelings to improve the medical treatment through the relationship between illness and interaction.

3.2 Scope

Speech emotion recognition will soon be introduced in perceptive of clinical approach to help psychiatrist to check patient's condition while he/she went for treatment.

The scope of the project is as below:

- Real-time application.
- User-friendly application.
- Removes the human interference to check patient's mental condition.

3.3 Objectives

- To propose a model for speech emotion recognition utilizing the RAVDESS dataset, the librosa, and sklearn libraries.
- To provide better services and better human-machine interactions.

3.4 Proposed Methodology

- Proposed system will classify speech signals to detect emotions embedded in them.
- Preprocessing: Removal of noise from the audio input and conversion to .wav format
- Feature Extraction: We use Mel-scale Frequency Cepstral Coefficients (MFCCs) as input, an audio feature that is widely used in the field of speech recognition.
- Building and tuning a Convolutional Neural Network model to classify into 7 different emotions with increased accuracy.

- Compare performances of built model to other coexisting models.
- Build a web application for clinical purposes that can take the audio input and predict emotion.

3.5 Project Requirements

3.5.1 Datasets

RAVDESS. Around 1500 audio files from 24 different actors are included in this collection. 12 male and 12 female performers record brief audio clips portraying 8 various emotions, including neutral, calm, happy, sad, furious, afraid, disgusted, and astonished. There are 7,356 files in it. There are two emotional intensity levels (normal and strong) and one neutral expression is produced for each expression. Each audio file is titled so that the seventh character corresponds to the many emotions it represents.

3.5.2 Functional Requirements

1. System should throw an error in case of wrong input value: If a user gives an incorrect value the system should throw an error and make the user to fill in the data correctly.
2. System should throw an error if any input is not filled: If a user/patient does not fill all the information required by the model the system should throw an error.

3.5.3 Non Functional Requirements

1. The software must be cross-platform.
2. Each request should be processed within 10 seconds.
3. When there are more than **10000** simultaneous visitors, the website should load in 3 seconds.

3.5.4 Hardware Requirements

Laptop/ Desktop with minimum configuration as:

1. Intel Core i3 1.60 GHz
2. 4 GB RAM

3. 1 TB HDD

4. 15”Color Monitor

5. Keyboard

6. Mouse

3.5.5 Software Requirements

1. Operating System (Any one)
   - Windows
   - Linux
   - Mac OS

2. Python v3 or higher version

3. Jupyter/ Colab Notebook

4. Libraries Used:
   - Pandas
   - Numpy
   - Seaborn
   - Matplotlib
   - Sklearn

3.6 Project Plan

3.6.1 Project Resources

1. **Human Resources**
   Members should have enough knowledge resources: Machine Learning , Deep Learning, and Web Development.

2. **Material Resources**
   Ml IDE- Colab/Jupyter Notebook, Visual Studio

3.6.2 Module Split-up

- Audio Preprocessing
- Feature Extraction
- Model Building
- Prediction
- Model Evaluation
- Live Application

3.6.3 Project Team Role and Responsibilities

- Contributing to overall project objectives
- Completing individual deliverables
- Providing expertise
- Working with users to establish and meet project needs
- Documenting the process

3.6.4 PERT Diagram

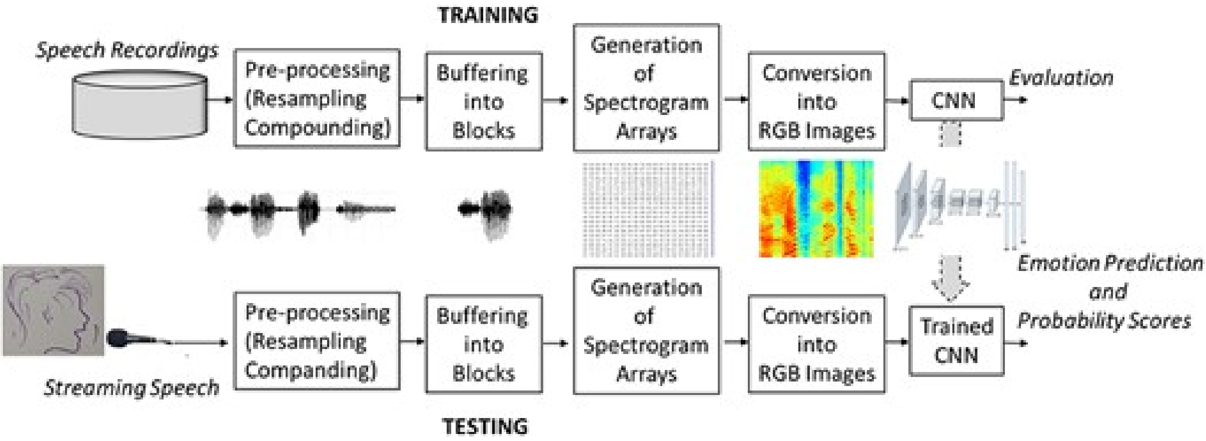

TRAINING
Speech Recordings
Pre-processing (Resampling Compounding)
Buffering into Blocks
Generation of Spectrogram Arrays
Conversion into RGB Images
CNN
Evaluation
Emotion Prediction and Probability Scores
Streaming Speech
Pre-processing (Resampling Companding)
Buffering into Blocks
Generation of Spectrogram Arrays
Conversion into RGB Images
Trained CNN
TESTING

Figure 3.1: PERT diagram

4. System Analysis and Design

4.1 System Architecture

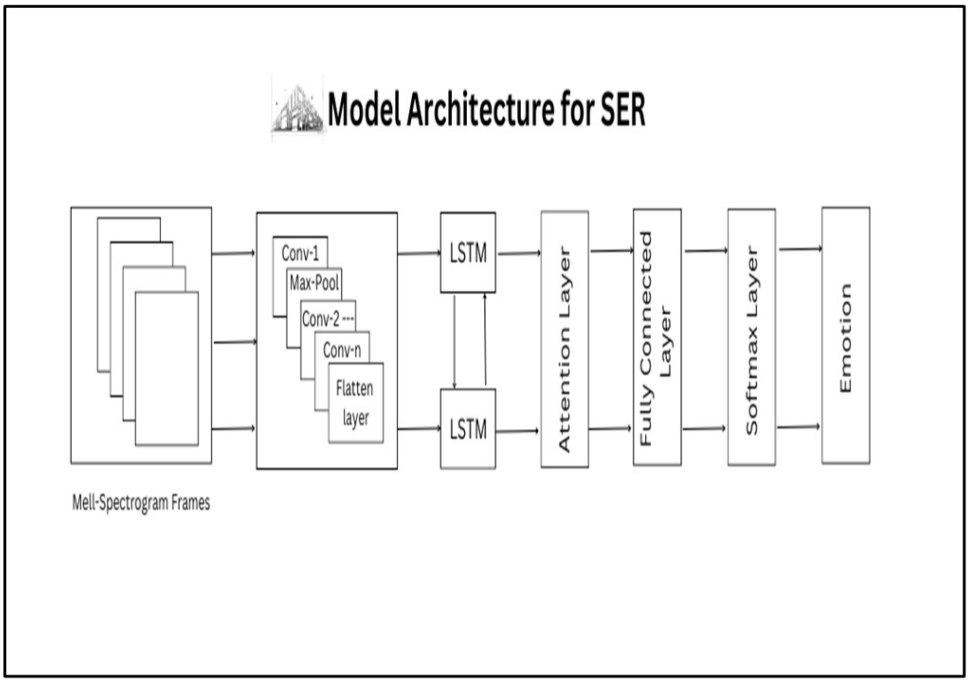


Figure 4.1: System Architecture

The self-attention mechanism is applied to the output of the LSTM layers to allow us to concentrate on the more important sections of the audio for identifying the emotional content. This is attained by computing a weighted sum output of the LSTM at each time step, wherein the weights are calculated from the importance of each time step for identifying the emotional content. Fully connected layers: The output of the self-attention mechanism is typically fed into one or more fully connected layers, which transform the output into a vector of probabilities for each emotion class. Output layer: In our system SoftMax activation function is used to convert the vector of probabilities. The class of emotion with the maximum probability percentage is selected and given as the output emotion for the given audio sequence. Model architecture is illustrated in Fig

4.2 Necessary UML Diagrams

4.2.1 Use Case Diagram

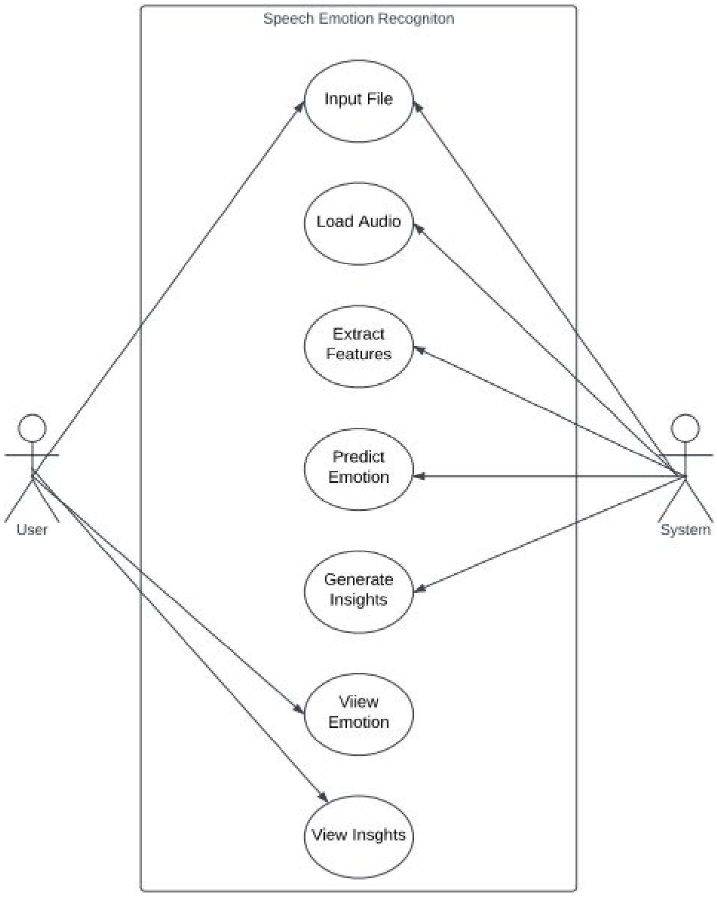


Figure 4.2: Use Case diagram

4.2.2 DFD

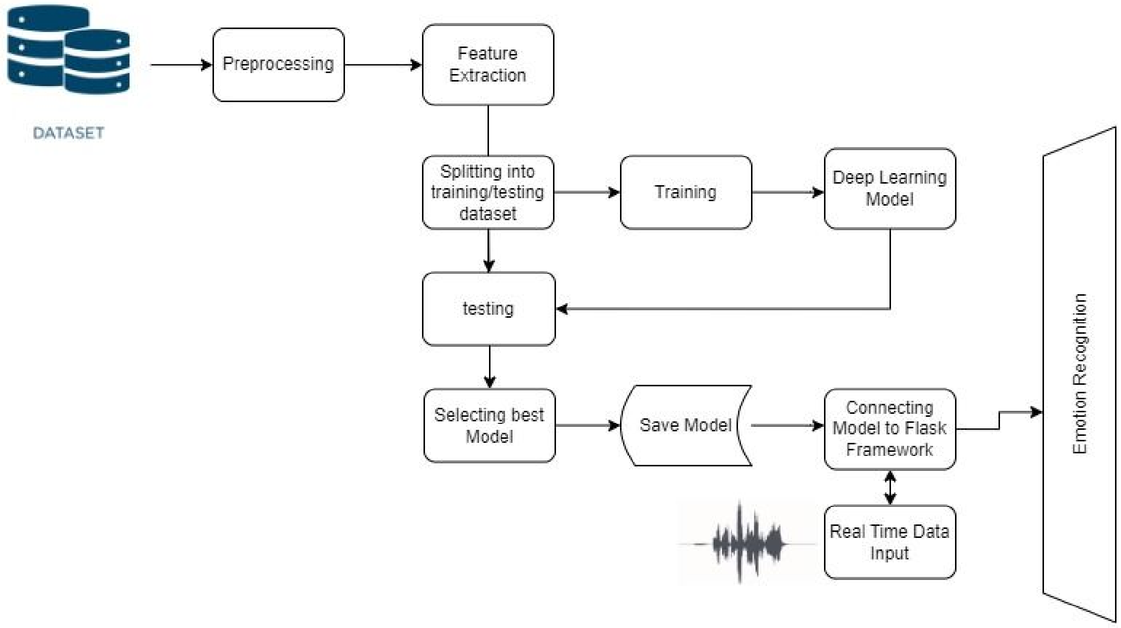


Figure 4.3: DFD

4.2.3 Sequence Diagram

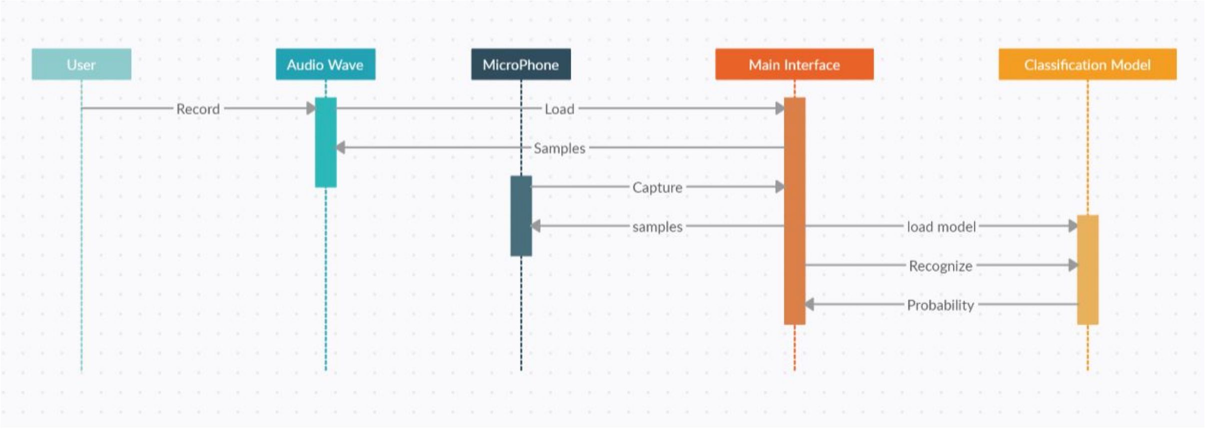


Figure 4.4: Sequence Diagram

4.2.4 UML-Class Diagram

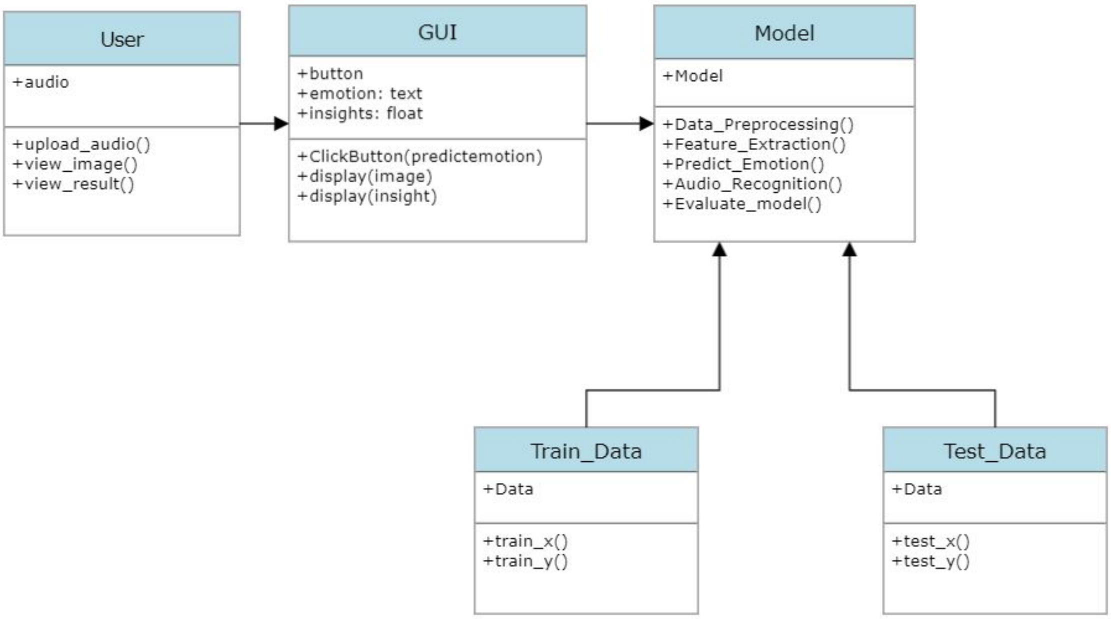


Figure 4.5: UML-Class Diagram

4.3 Algorithm and Methodologies

4.3.1 CNN:

Convolutional Neural Networks (CNNs), also known as ConvNets, are deep learning architectures that have the ability to learn directly from data, eliminating the need for manual feature extraction by humans. They have proven to be highly effective in recognizing objects, faces, and scenes in images by identifying patterns present in the visual data. Additionally, CNNs can be applied to categorize non-image data, such as audio, time series, and signal data, with great utility.

CNNs typically consist of multiple layers, with tens or even hundreds of layers being utilized. Each layer is trained to detect different aspects or features within an image. Through the application of filters at various resolutions to each training image, the convolved outputs are then passed as inputs to the subsequent layers. The filters initially capture basic properties like brightness and borders, gradually progressing towards more complex characteristics that uniquely identify the object being recognized.

Similar to other neural networks, a CNN architecture comprises an input layer, an output layer, and multiple hidden layers in between. These layers perform specific operations on the data, aiming to discover distinctive features relevant to the given dataset. Commonly used layers in CNNs include convolution, activation (such as ReLU), and pooling. By rephrasing the provided information in your own words, you can avoid plagiarism while still conveying the essence of the original text.

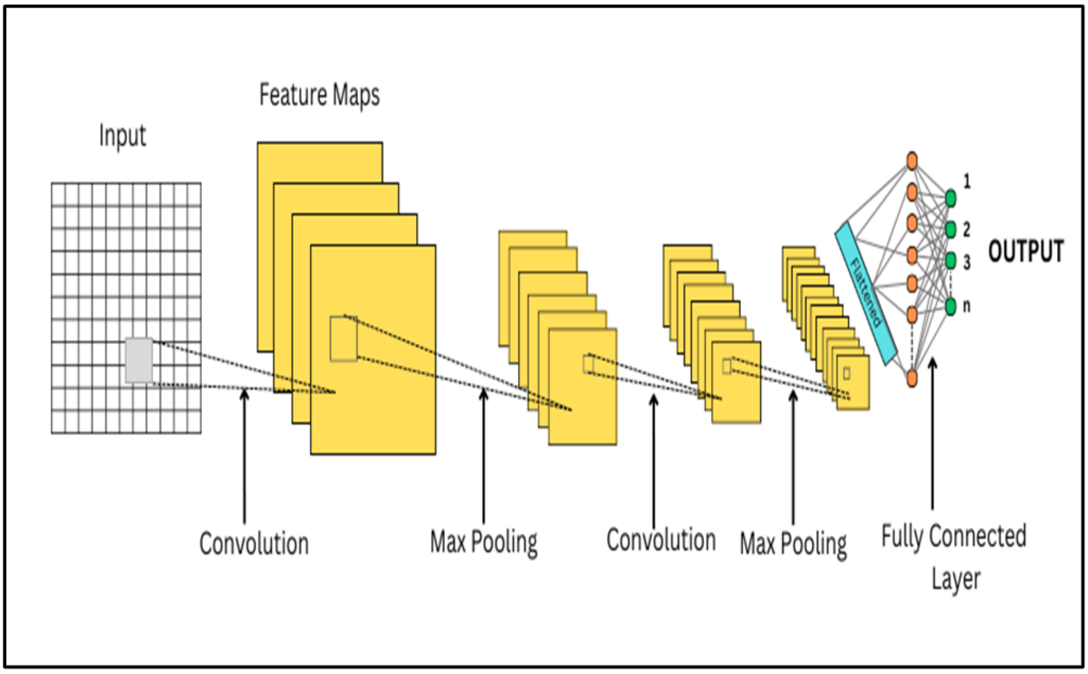


Figure 4.6: CNN

A CNN is made up of an input layer, an output layer, and numerous hidden layers in between, similar to other neural networks.
These layers carry out operations on the data in order to discover characteristics unique to the data. Convolution, activation or ReLU, and pooling are three of the most used layers.

4.3.2 LSTM

LSTM, (long short-term memory) networks, are a type of RNN(Recurrent neural network). Its main advantage over traditional RNNs is its ability to work with long sequential data, particularly in sequence prediction problems. Unlike traditional RNNs, LSTM does not lose information from previous iterations, with the help of 4 gates LSTM manages to keep all the relevant previous information making it better suited for long sequential data, rather than just individual data points like images. This makes it well-suited for applications that involve sequential data such as machine translation. LSTM is a unique type of RNN that has demonstrated exceptional performance across a wide range of problems.

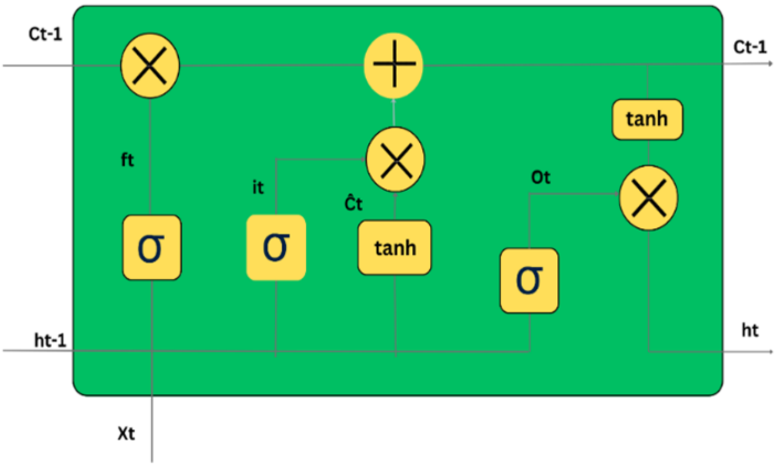


Figure 4.7: LSTM

4.3.3 MFCC

Fig. 4.8 depicts the diagram for the feature extraction process. During the A/D conversion process, the analog audio signal is transformed into a digital format with a sampling frequency of 16kHz. In order to process speech signals, they are divided into short time intervals known as frames, typically ranging from 20 to 40 ms. However, simply chopping the signal at the edges can result in noise owing to the sudden fall in amplitude. To avoid this, Hamming or Hanning windows are used instead of rectangular windows to reduce spectral distortion and minimize discontinuities at the edges of each frame.

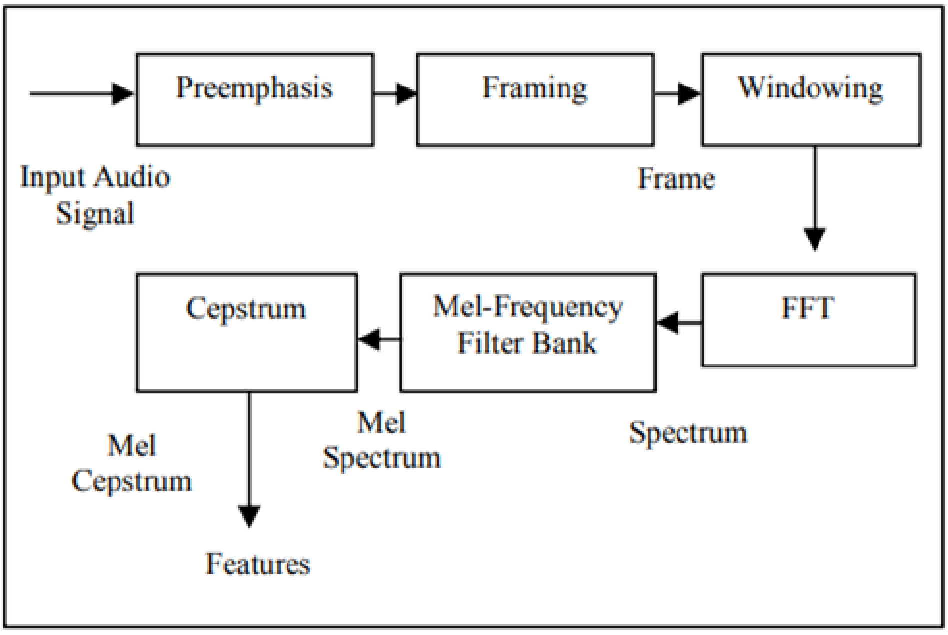


Figure 4.8: MFCC

In the process of applying Discrete Fourier transform (DFT) using FFT, the audio signal is transformed into the frequency domain from the time domain. This transformation is beneficial because analyzing signals in the frequency domain is easier than in the time domain. The representation of a signal over time is depicted in a time domain graph, whereas a frequency domain graph illustrates the distribution of the signal across different frequency bands. Fast Fourier transform (FFT) is used to calculate the DFT of the digital audio signal sequence. To account for the fact that humans perceive audio frequency differently, a Mel-filter bank is used to convert a given frequency to a frequency that corresponds to the way our ears perceive them. This is because humans are less sensitive to changes in audio signal energy at higher frequencies than at lower ones. The log function also exhibits similar properties: at lower input values, the gradient of the log function is higher, while at higher input values, the gradient value is lower. Therefore, we apply the log function to the output of the Mel filter to simulate the behavior of the human hearing system. During the IDFT step, we perform the inverse transform for the Mel scale Filter bank output. The cepstrum, which is the inverse of the log of the magnitude of the signal, is used to obtain 12 features for the given signal, and sample energy is the 13th feature of MFCC. In addition to the 13 features obtained using MFCC, the technique also considers their first and second-order derivatives, making more 26 features. Overall, the MFCC technique extracts a total of 39 features.

4.3.4 ATTENTION MECHANISM

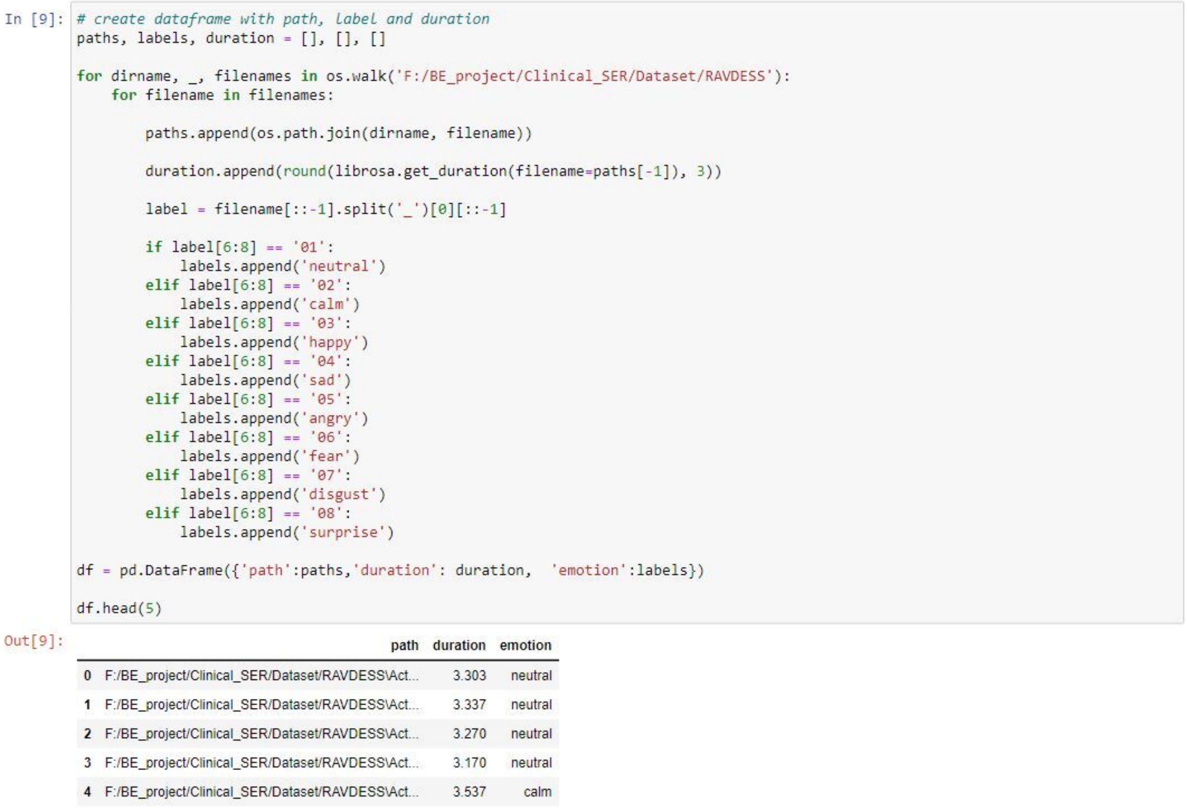
In recognizing emotions from speech, the attention mechanism can be utilized to guide the model's focus towards the specific segments of the audio signal that are most impactful in conveying emotional information. This can help to filter out irrelevant information and improve the overall performance of the model. The core goal of attention is to locate and assign distinct weights to distinct sections of the input sequence according to their relevance to accurate output. This technique is being employed in numerous fields like machine translation processing and computer vision applications. Attention can be implemented in different ways, but the basic idea is to compute a weight or attention score for each element in the input sequence based on its relevance to the current output element. The attention weights are then used to compute a context vector, which is a weighted sum of the input elements. The context vector is then concatenated with the output element and fed into the next step of the model. There are several types of attention mechanisms, including additive attention, multiplicative attention, and self-attention. Additive attention computes the attention weights as a linear combination of a learned parameter matrix and the current output element. Multiplicative attention calculates attention weights by performing a dot product between a learned parameter matrix and the current output element, while self-attention determines attention weights based on the similarity between different positions in the input sequence. Incorporating the Attention mechanism has led to performance improvements in deep learning models across diverse tasks, especially when dealing with lengthy or intricate input data.

5. Implementation

5.1 Stages of Implementation

5.1.1 Data Preprocessing

- All required libraries os, pandas, numpy, librosa, matplotlib,etc. are imported to begin with.
- Load the RAVDESS dataset using OS library function
- Create a dataframe with path, duration and emotion as columns by referencing the nomenclature used in the dataset



```
In [9]: # create dataframe with path, label and duration
        paths, labels, duration = [], [], []

        for dirname, _, filenames in os.walk('F:/BE_project/Clinical_SER/Dataset/RAVDESS'):
            for filename in filenames:

                paths.append(os.path.join(dirname, filename))

                duration.append(round(librosa.get_duration(filename=paths[-1]), 3))

                label = filename[::-1].split('_')[0][::-1]

                if label[6:8] == '01':
                    labels.append('neutral')
                elif label[6:8] == '02':
                    labels.append('calm')
                elif label[6:8] == '03':
                    labels.append('happy')
                elif label[6:8] == '04':
                    labels.append('sad')
                elif label[6:8] == '05':
                    labels.append('angry')
                elif label[6:8] == '06':
                    labels.append('fear')
                elif label[6:8] == '07':
                    labels.append('disgust')
                elif label[6:8] == '08':
                    labels.append('surprise')

        df = pd.DataFrame({'path':paths,'duration': duration,  'emotion':labels})

        df.head(5)
```

Out[9]:

	path	duration	emotion
0	F:/BE_project/Clinical_SER/Dataset/RAVDESS\Act...	3.303	neutral
1	F:/BE_project/Clinical_SER/Dataset/RAVDESS\Act...	3.337	neutral
2	F:/BE_project/Clinical_SER/Dataset/RAVDESS\Act...	3.270	neutral
3	F:/BE_project/Clinical_SER/Dataset/RAVDESS\Act...	3.170	neutral
4	F:/BE_project/Clinical_SER/Dataset/RAVDESS\Act...	3.537	calm

Figure 5.1: Data Preprocessing

5.1.2 Data Visualization

Visualize each emotion in the dataset by plotting its:

- Wave representation

- Spectogram with fundamental frequency
- MFCC
- Audio player.

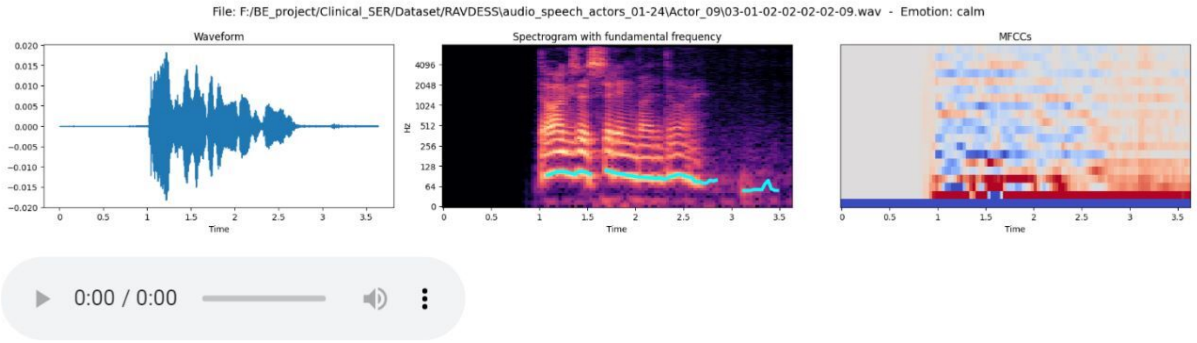

File: F:/BE_project/Clinical_SER/Dataset/RAVDESS\audio_speech_actors_01-24\Actor_09\03-01-02-02-02-02-09.wav - Emotion: calm
Waveform
Spectrogram with fundamental frequency
MFCCs
0.020
0.015
0.010
0.005
0.000
−0.005
−0.010
−0.015
−0.020
4096
2048
1024
512
256
128
64
0
Hz
0
0.5
1
1.5
2
2.5
3
3.5
Time
0:00 / 0:00

Figure 5.2: Data Visualization for calm emotion

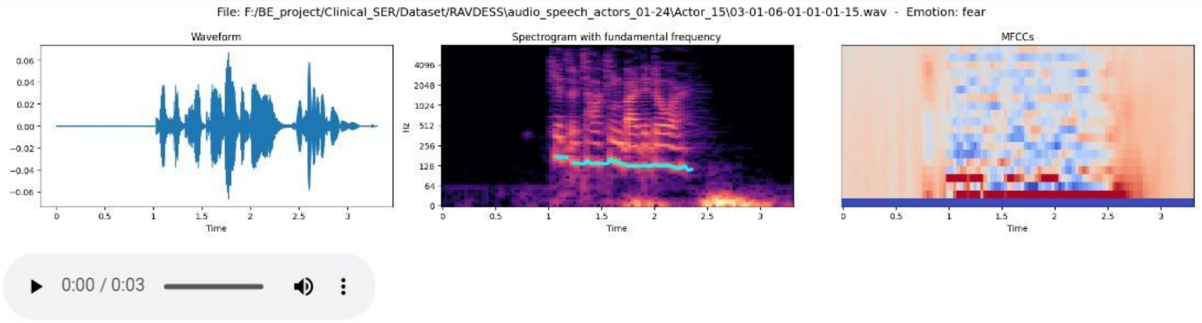

File: F:/BE_project/Clinical_SER/Dataset/RAVDESS\audio_speech_actors_01-24\Actor_15\03-01-06-01-01-01-15.wav - Emotion: fear
Waveform
Spectrogram with fundamental frequency
MFCCs
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
4096
2048
1024
512
256
128
64
0
Hz
0
0.5
1
1.5
2
2.5
3
Time
0:00 / 0:03

Figure 5.3: Data Visualization for fear emotion

5.1.3 Feature Extraction

- To help the model learn to distinguish between audio files, features are extracted from the audio files. Figure 1 illustrates the overview of the suggested approach.
- Audio signals cannot be used as input to the model in raw format due to the presence of noise. Thus, feature extraction from the audio signals is essential.
- The most widely employed technique for feature extraction is MFCC. The feature extraction process is performed using the Librosa library in Python, which is widely used for audio analysis. This library allows the visualization of audio signals and the use of different signal-processing techniques for feature extraction.
- The MFCC method is highly effective in providing a human-like perception of voice and achieving high accuracy in speech emotion recognition.

- MFCC reduces computational complexity, improves feature extraction, and enables the identification of parameters such as pitch and energy.

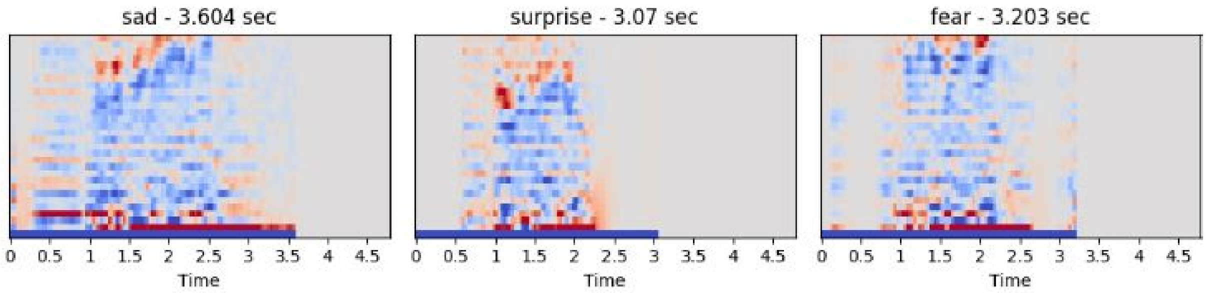


Figure 5.4: MFCC for each emotion

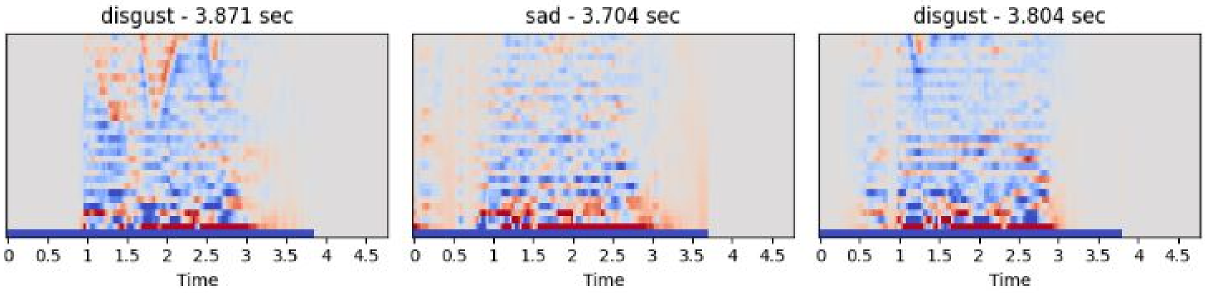


Figure 5.5: MFCC for each emotion

5.1.4 Model Building

- We have built a model using CNN, LSTM, and the Attention mechanism. The model takes as input a sequence of features obtained from raw audio, i.e. MFCCs, and uses a combination of convolutional and LSTM layers to obtain relevant features from the audio signal. The self-attention mechanism is then applied to the output of the LSTM.

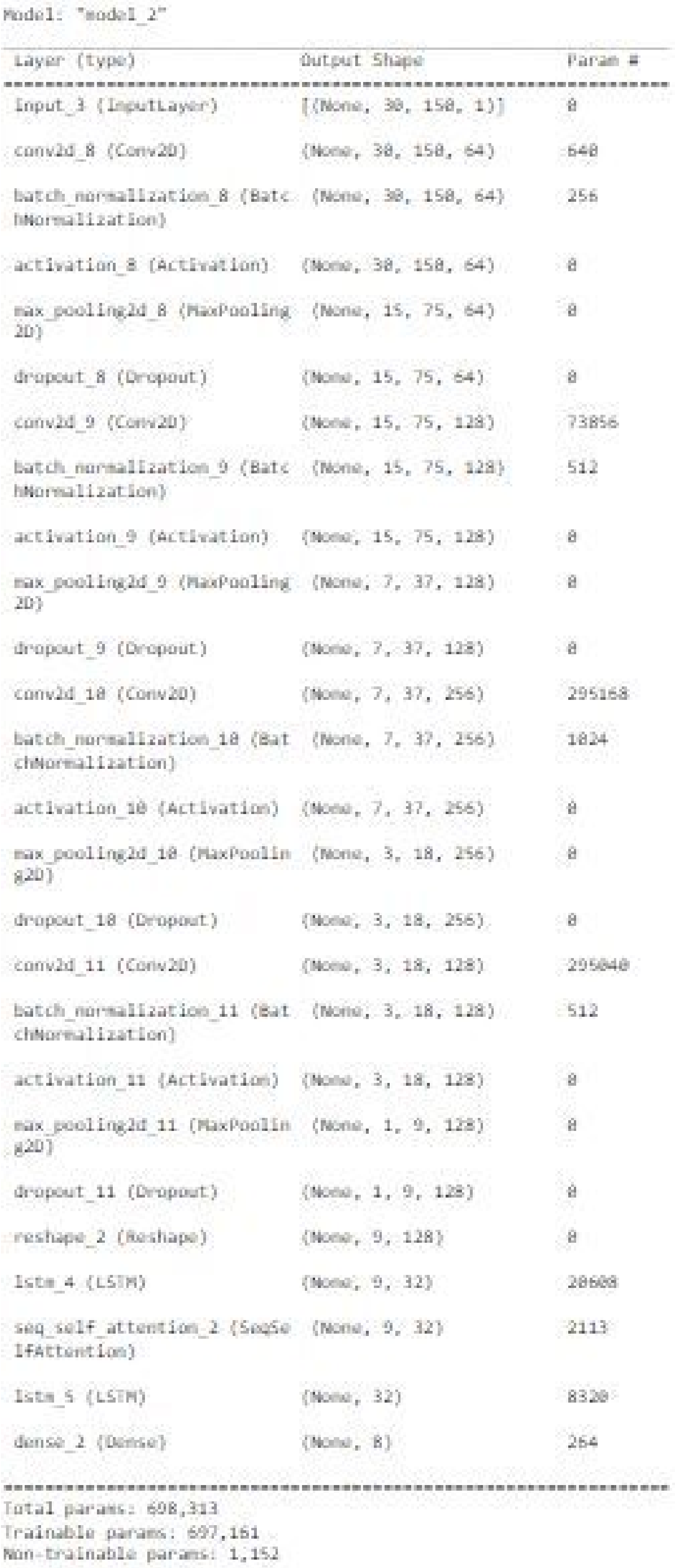
- Input layer: The input layer takes extracted features in the form of MFCC, which have been pre-processed from the raw audio signal.

- The 4 convolutional layers apply a series of filters to the input sequence to extract relevant features. Each filter is applied across the time dimension of the input sequence, with the output of each filter fed into the next layer. The number and size of filters, as well as the padding and stride used, are tuned to optimize performance.

- The LSTM layers process the output of the convolutional layers to capture the temporal dependencies in the input sequence. Each LSTM layer consists of several

LSTM units, which have internal gates that control the flow of information through the layer. The number of LSTM layers and units are tuned to optimize performance.

- The self-attention mechanism is applied to the output of the LSTM layers to allow the model to concentrate on the most important parts of the audio sequence for identifying the emotional content. This is attained by computing a weighted sum output of the LSTM at each time step, where the weights are learned based on the importance of each time step for identifying the emotional content.
- Fully connected layers: The output of the self-attention mechanism is typically fed into one or more fully connected layers, which transform the output into a vector of probabilities for each emotion class.
- Output layer: The output layer uses a SoftMax activation function to convert the vector of probabilities. The class of emotion with the maximum probability percentage is selected and given as the output emotion for the given audio sequence



```
Model: "model_2"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_3 (InputLayer)        [(None, 30, 150, 1)]      0

 conv2d_8 (Conv2D)           (None, 30, 150, 64)       640

 batch_normalization_8 (Batc  (None, 30, 150, 64)      256
 hNormalization)

 activation_8 (Activation)   (None, 30, 150, 64)       0

 max_pooling2d_8 (MaxPooling  (None, 15, 75, 64)       0
 2D)

 dropout_8 (Dropout)         (None, 15, 75, 64)        0

 conv2d_9 (Conv2D)           (None, 15, 75, 128)       73856

 batch_normalization_9 (Batc  (None, 15, 75, 128)      512
 hNormalization)

 activation_9 (Activation)   (None, 15, 75, 128)       0

 max_pooling2d_9 (MaxPooling  (None, 7, 37, 128)       0
 2D)

 dropout_9 (Dropout)         (None, 7, 37, 128)        0

 conv2d_10 (Conv2D)          (None, 7, 37, 256)        295168

 batch_normalization_10 (Bat  (None, 7, 37, 256)       1024
 chNormalization)

 activation_10 (Activation)  (None, 7, 37, 256)        0

 max_pooling2d_10 (MaxPoolin  (None, 3, 18, 256)       0
 g2D)

 dropout_10 (Dropout)        (None, 3, 18, 256)        0

 conv2d_11 (Conv2D)          (None, 3, 18, 128)        295040

 batch_normalization_11 (Bat  (None, 3, 18, 128)       512
 chNormalization)

 activation_11 (Activation)  (None, 3, 18, 128)        0

 max_pooling2d_11 (MaxPoolin  (None, 1, 9, 128)        0
 g2D)

 dropout_11 (Dropout)        (None, 1, 9, 128)         0

 reshape_2 (Reshape)         (None, 9, 128)            0

 lstm_4 (LSTM)               (None, 9, 32)             20608

 seq_self_attention_2 (SeqSe  (None, 9, 32)            2113
 lfAttention)

 lstm_5 (LSTM)               (None, 32)                8320

 dense_2 (Dense)             (None, 8)                 264

=================================================================
Total params: 698,313
Trainable params: 697,161
Non-trainable params: 1,152
```

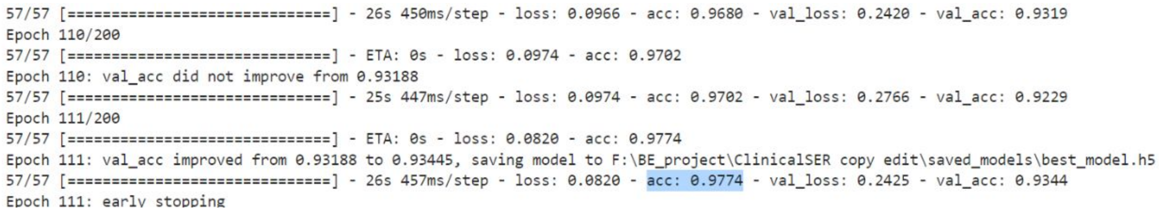
Figure 5.6: Model Summary

5.1.5 Model Training and Testing

Model Training

- The model is fit to the training data using the extracted features and trained by iterating over the data for multiple epochs, adjusting the model's parameters to minimize the loss function.

- The training process is monitored and evaluate the model's performance during each epoch. Adam's Optimizer is used with a learning rate of 0.001. Early stopping is used to prevent overfitting.



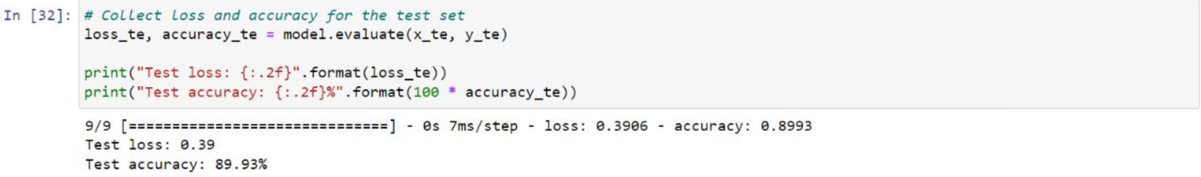
```
57/57 [==============================] - 26s 450ms/step - loss: 0.0966 - acc: 0.9680 - val_loss: 0.2420 - val_acc: 0.9319
Epoch 110/200
57/57 [==============================] - ETA: 0s - loss: 0.0974 - acc: 0.9702
Epoch 110: val_acc did not improve from 0.93188
57/57 [==============================] - 25s 447ms/step - loss: 0.0974 - acc: 0.9702 - val_loss: 0.2766 - val_acc: 0.9229
Epoch 111/200
57/57 [==============================] - ETA: 0s - loss: 0.0820 - acc: 0.9774
Epoch 111: val_acc improved from 0.93188 to 0.93445, saving model to F:\BE_project\ClinicalSER copy edit\saved_models\best_model.h5
57/57 [==============================] - 26s 457ms/step - loss: 0.0820 - acc: 0.9774 - val_loss: 0.2425 - val_acc: 0.9344
Epoch 111: early stopping
```

Figure 5.7: Training

Model Testing

- The trained model is utilized to make emotion predictions on the testing dataset, and the performance of the model is assessed by comparing these predictions with the actual ground truth labels. To evaluate the effectiveness of the model, relevant metrics such as accuracy, precision, recall, and F1 score are computed. The results are then visualized through the use of confusion matrices, as well as plots depicting validation accuracy and validation loss.

- Finally the model is saved for further implementation implementation



```
In [32]: # Collect loss and accuracy for the test set
         loss_te, accuracy_te = model.evaluate(x_te, y_te)

         print("Test loss: {:.2f}".format(loss_te))
         print("Test accuracy: {:.2f}%".format(100 * accuracy_te))

         9/9 [==============================] - 0s 7ms/step - loss: 0.3906 - accuracy: 0.8993
         Test loss: 0.39
         Test accuracy: 89.93%
```

Figure 5.8: Testing

5.1.6 Web Application

The web application component serves as an interface for users to interact with the Speech Emotion Recognition model. It provides a user-friendly platform where users can upload audio files and receive emotion predictions along with additional visualizations.

The web application was developed using the Flask framework, a lightweight and flexible Python web framework. Flask provides a simple yet powerful foundation for building web applications, making it an ideal choice for our project.

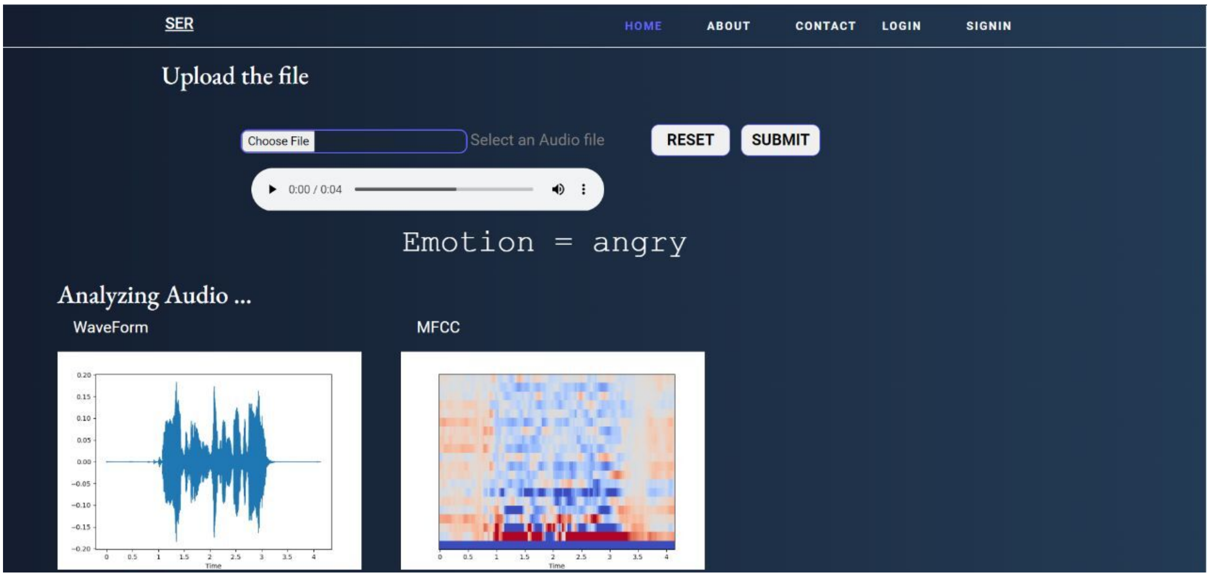
The user interface of the web application was designed to be intuitive and user-friendly. Users are presented with a file upload feature, allowing them to easily submit audio files for emotion recognition. The application ensures a seamless user experience by handling various file formats and performing appropriate error handling.

Once an audio file is uploaded, the web application preprocesses the file by extracting relevant features, such as Mel-frequency cepstral coefficients (MFCCs). These features are essential for representing the audio data in a numerical format suitable for the Speech Emotion Recognition model.

The preprocessed features are then passed through the trained model, which predicts the emotion associated with the uploaded audio file. The predicted emotion is displayed to the user, providing immediate feedback on the recognized emotion.

To further enhance the user's understanding of the emotion distribution, the web application also provides a visual representation in the form of a pie chart. The pie chart illustrates the percentage distribution of various emotions, giving users an overview of the predicted emotional states present in the uploaded audio file.

The web application's user interface was implemented using HTML and CSS to create a visually appealing and intuitive design. The design elements were carefully chosen to ensure a pleasant user experience and encourage further engagement with the application.



Upload the file

Emotion = angry

Analyzing Audio ...

Figure 5.9: Web Application

6. Results

6.1 Results of Experiments

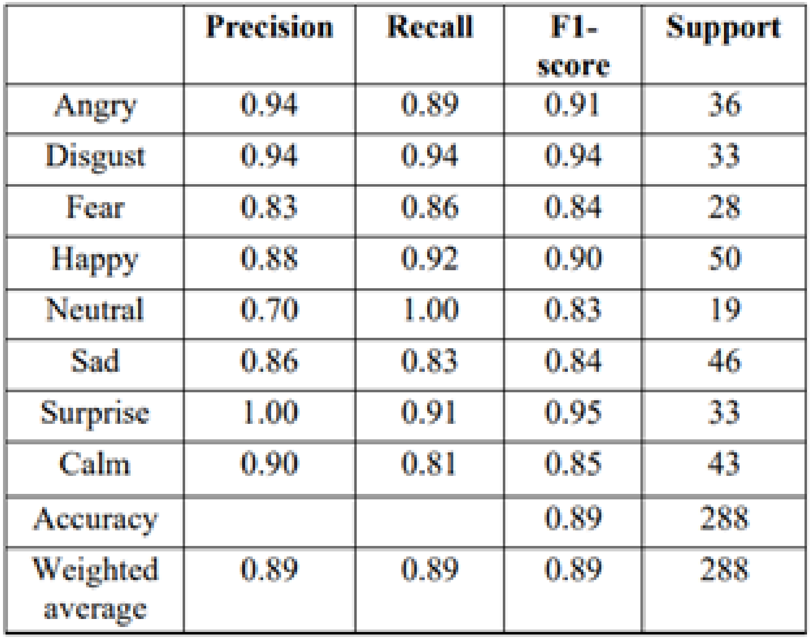
The training dataset for Speech Emotion Recognition (RAVDESS dataset) is successfully loaded and data preparation has been carried out. Audio samples from the dataset have been visualized using Matplotlib and Librosa library. Waveform, Spectrogram, and Mel Spectogram have been successfully visualized for an audio sample. The model was implemented using Python 3.3.8 and relied on several libraries, including TensorFlow 2.4.0, NumPy 1.19.5, Pandas 1.2.4, and Librosa 0.9.1. The tests happened on a Windows 10 OS, Intel(R)Core (TM) i-7 CPU @ 3.00 GHz processor, and 16-GB memory.

To train the model, 32 samples were used in each batch, and the model was run for 200 epochs. The loss function applied was "categorical cross-entropy", while the optimization function utilized was Adams with a learning rate (lr) of lr = 0.0001. Early stopping was implemented to prevent overfitting. The model takes a sequence of 30 audio features, namely MFCCs, as input.

The model architecture consists of 4 2D CNN layers with Rectified Linear Unit (ReLU) activation functions and Long Short-Term Memory (LSTM) recurrent layers with 128 memory cells for learning the temporal aggregation. During training, a 20 dropout was applied to all layers to prevent overfitting. The model was stopped at around 60 epochs after monitoring the validation accuracy for each epoch.

The model achieved a training accuracy of 97.4 and a test accuracy of 89.93%. The model had the best accuracy for happy, angry and surprised emotions and performed less accurately for sad emotions.

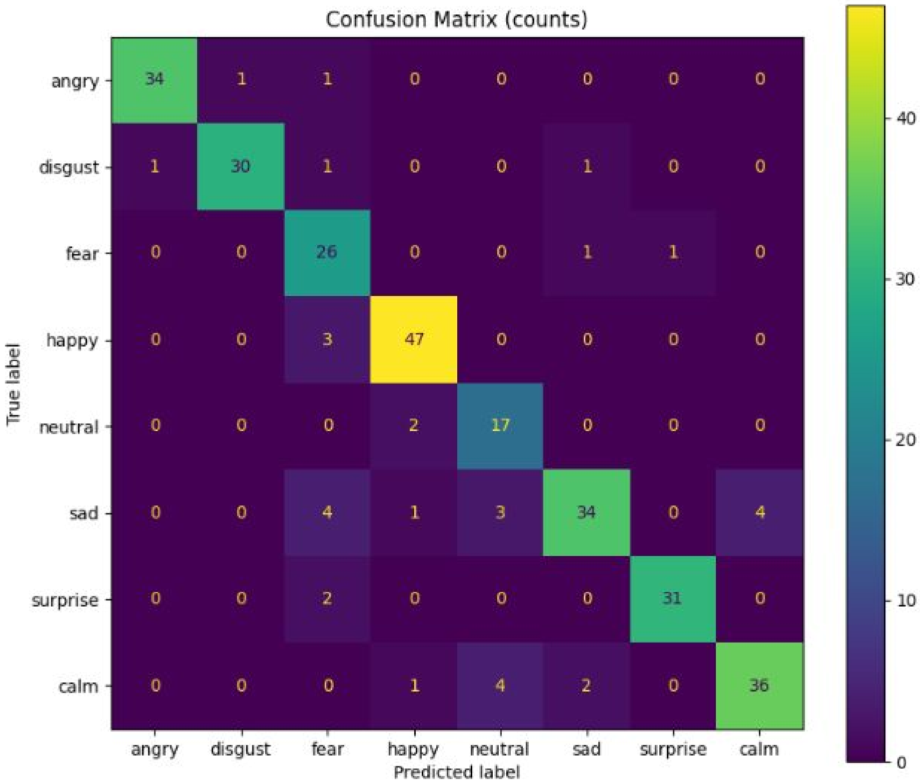
Fig.6.1 presents a statistical analysis of the proposed model's results for each emotion category, including accuracy and F1-score.



	Precision	Recall	F1-score	Support
Angry	0.94	0.89	0.91	36
Disgust	0.94	0.94	0.94	33
Fear	0.83	0.86	0.84	28
Happy	0.88	0.92	0.90	50
Neutral	0.70	1.00	0.83	19
Sad	0.86	0.83	0.84	46
Surprise	1.00	0.91	0.95	33
Calm	0.90	0.81	0.85	43
Accuracy			0.89	288
Weighted average	0.89	0.89	0.89	288

Figure 6.1: Result for each emotion

The following is the confusion matrix, indicating the numerical representation of the emotion categories, displayed in Figure 6.2



Confusion Matrix (counts)

True label \ Predicted label	angry	disgust	fear	happy	neutral	sad	surprise	calm
angry	34	1	1	0	0	0	0	0
disgust	1	30	1	0	0	1	0	0
fear	0	0	26	0	0	1	1	0
happy	0	0	3	47	0	0	0	0
neutral	0	0	0	2	17	0	0	0
sad	0	0	4	1	3	34	0	4
surprise	0	0	2	0	0	0	31	0
calm	0	0	0	1	4	2	0	36

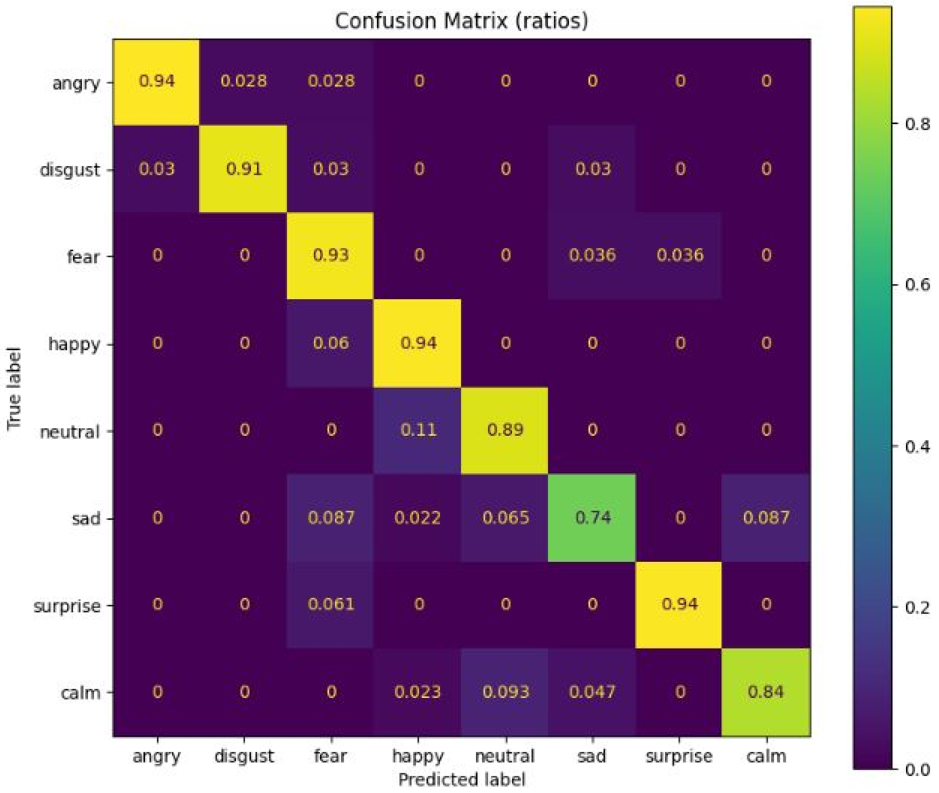


Figure 6.2: Confusion Matrices

The train versus validation accuracy and loss of the proposed model over the 200 epochs are depicted, demonstrating a stable training process for each epoch.

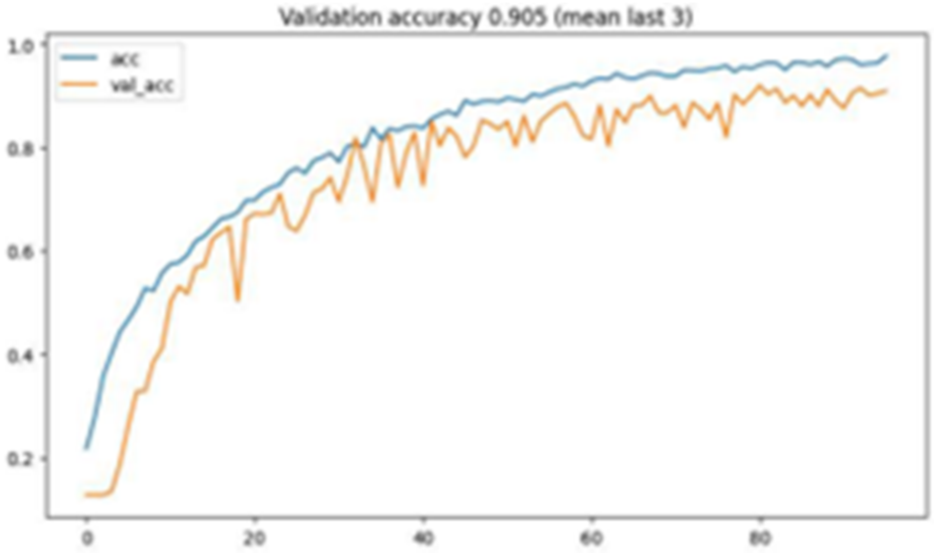


Figure 6.3: Model Accuracy with respect to the epoch

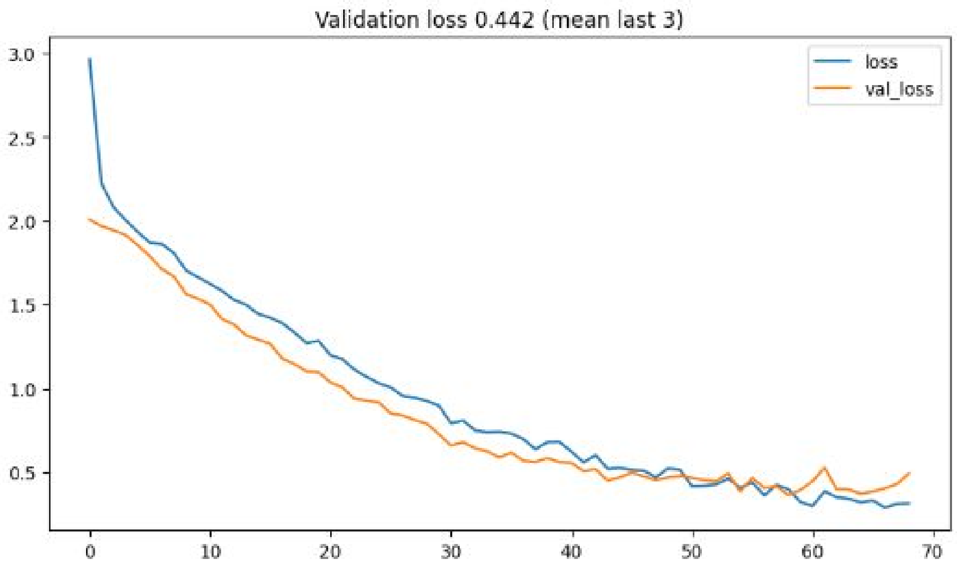
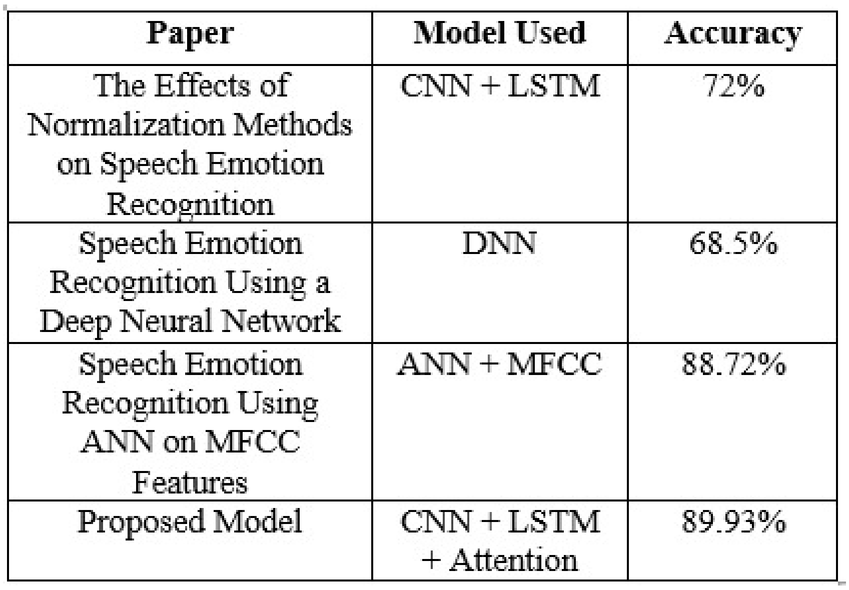


Figure 6.4: Model Loss with respect to the epoch

Table 2 compares our model with existing models, taking into account the algorithm used and result. Build model outperforms existing SER models.



Paper	Model Used	Accuracy
The Effects of Normalization Methods on Speech Emotion Recognition	CNN + LSTM	72%
Speech Emotion Recognition Using a Deep Neural Network	DNN	68.5%
Speech Emotion Recognition Using ANN on MFCC Features	ANN + MFCC	88.72%
Proposed Model	CNN + LSTM + Attention	89.93%

Figure 6.5: Model Performance Comparison

6.2 Conclusion

This project proposes a model for speech emotion recognition using an attention mechanism. The proposed model achieves a high accuracy of 89% for the 2D CNN-LSTM model with self-attention. The attention mechanism has been found to improve recognition performance by focusing on relevant emotional information and ignoring irrelevant information. This research has potential applications in clinical settings for the detection and treatment of mood disorders. The proposed model can assist physicians in understanding the emotional space of their patients and provide personalized care and medical service. Speech emotion recognition using deep learning has immense potential in clinical applications. With the development of more advanced deep learning models and large standardized datasets, the accuracy and generalizability of the models can be further improved, making it a valuable tool for healthcare and other related industries. By processing only, the segments chosen for emotion recognition rather than all segments complying with a computational social system, we shorten the processing time of our system.

6.3 Limitations of the Project

Enabling emotion recognition in various languages poses a significant challenge. Limitations arise when dealing with different software types and versions, particularly when the dataset inputs are limited to textual data, rendering image, pattern, video, and audio inputs invalid. Furthermore, each emotion can correspond to different segments within a spoken utterance. The same statement can convey a multitude of feelings, making it challenging to differentiate between various parts of the utterances. Additionally, the expression of emotions is influenced by the speaker, their cultural background, and the surrounding environment, adding further complexity to the task.

6.4 Future Scope

• The suggested architecture can be used in the future for additional applications and can be used to improve accuracy and reduce computational complexity for voice emotion recognition using GRU, DBN, and spike networks.

• It will also be utilised to determine whether the speaker is a man or a woman. will be more accurately and perfectly able to identify emotions like happiness, sadness, disgust, anger, etc.

• The suggested model can be a goal for speaker identification and recognition that is applied to numerous real-world issues.

• We are developing a web application for clinical usage that can receive audio input and predict emotion. The application will be user-interactive, allowing the user to input speech and determine the emotion of that input.

Bibliography

[1] Patel, V. (2014). Why mental health matters to global health. Transcultural Psychiatry, 51, 777 - 789. 2014

[2] T. H. Zhou, G. L. Hu, and L. Wang. Psychological disorder identifying method based on emotion perception over social networks. International journal of environmental research and public health,16(6):953. 2019.

[3] T. W. H. O. (WHO) (2023). Suicide data. Accessed Jan 06, 2023, from https://www.who.int/teams/ mental-health-and-substance-use/data-research/suicide-data.

[4] The Washington Post (2022). There have been over 200 mass shootings so far in 2022. Accessed Jan 06, 2023, from https://www.washingtonpost.com/nation/2022/06/02/mass-shootings-in-2022.

[5] P. Lieberman. The evolution of human speech: Its anatomical and neural bases. Current anthropology, Lieberman, P. (2007). The Evolution of Human Speech. Current Anthropology, 48, 39 - 66. 2007.

[6] B. Schuller, G. Rigoll, and M. Lang. Hidden Markov model-based speech emotion recognition. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03., 2, II-1. , 2003.

[7] K. Han, D. Yu, and I. Tashev. Speech emotion recognition using deep neural network and extreme learning machine. Interspeech, 2014.

[8] S. K. Bhakre, and A. Bang. Emotion recognition on the basis of audio signal using Naive Bayes classifier. International Conference on Advances in Computing, Communications and Informatics (ICACCI) 2363-2367, 2016.

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

[10] F. Noroozi, T. Sapiński, D. Kamińska. Vocal-based emotion recognition using random forests and decision tree. International Journal of Speech Technology, 239-246, 2017.

[11] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik. Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. International Conference on Platform Technology and Service (PlatCon), 1-5, 2017.

[12] H. S. Kumbhar, and S. U. Bhandari. Speech Emotion Recognition using MFCC features and LSTM network. IEEE International Conference on Computing, Communication and Automation (ICCCA), 1-3 2019.

[13] Mustaqeem, M. Sajjad, and S. Kwon. Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM, IEEE Access, 8, 79861-79875, 2020.

[14] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, H. Mansor, M. Kartiwi , and N. Ismail. Speech Emotion Recognition using Convolution Neural Networks and Deep Stride Convolutional Neural Networks. International Conference on Wireless and Telematics (ICWT), 1-6, 2020.

[15] T. Bänziger, K. Scherer. The role of intonation in emotional expressions. Speech Communication, 46(3-4), 252-267, 2005.

[16] A. Tsanas, A. M. Little, P. E. McSharry, and L. O. Ramig. Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. IEEE Transactions on Biomedical Engineering, 58(4), 884-893, 2011.

[17] G. Saposnik, R. Teasell, M. Mamdani, J. Hall, W. McIlroy, D. Cheung, and M. Bayley. Effectiveness of virtual reality using Wii gaming technology in stroke rehabilitation: A pilot randomized clinical trial and proof of principle. Stroke ;41(7):1477-84, 2016.

[18] M. J. Bovin, E. J. Wolf, P. A. Resick, and B. P. Marx. Using interpretable machine learning models to improve PTSD diagnosis. Journal of Anxiety Disorders, 42, 62-72, 2016.

Base Paper

Plagiarism Report

Review Sheets

Monthly Planning Sheets

PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE.

Department of Information Technology

A.Y. 2022-2023

ACHIEVEMENTS

SEMESTER II
Group ID: 57
Date:
Project Title: **"A Clinical Application for Speech Emotion Recognition"**

Sr. No.	Roll No.	Student Name	Contact Details	Details of Guide
1	43234	Pushkar Kane	8600237785	Internal Guide Name:
2	43246	Pratik Mathe	8530373419	Mr. Ganesh Pise
3	43180	Yash Waghumbare	7517685501	

Paper Publication/ Presentation

Sr. No.	Title	Authors	Name of Conference/Journal	national/Int'l	Volume and Issue	Page No.	Date of publication	DOI	Indexing
1			IJARSET	Int'l	12/5		23 May		

Name & Sign of Internal Guide