

# **IMPLEMENTATION OF DEEP LEARNING MODEL FOR AUDIO AND VIDEO ANALYSIS OF HUMAN EMOTIONS.**

*Submitted in partial fulfillment of the requirements for the degree of*

## **Bachelor of Technology in Computer Science Engineering**

*by*

**Abishek Prakash-19BDS0161**

**Mallikarjun Hatti-19BCE0889**

**Thukila C-19BCT0247**

**Under the guidance of**

**Prof. Sathya K**

**SCOPE**

**VIT, Vellore.**



May,2023

## **DECLARATION**

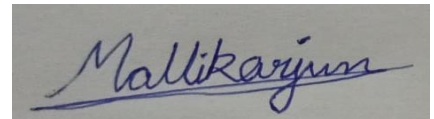
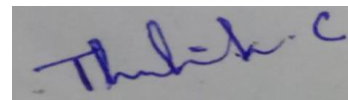
I hereby declare that the thesis entitled “**Implementation of Deep Learning Model for Audio And Video Analysis of Human Emotions**” submitted by us, for the award of the degree of Bachelor of Technology in Computer Science Engineering to VIT, is a record of bonafide work carried out by me under the supervision of **Prof.Sathya K.**

We further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Vellore

Date:

**Signature of the Candidate**

Handwritten signature of Mallikarjun in blue ink.Handwritten signature of Thibek C in blue ink.Handwritten signature of abishkek in blue ink.

## **CERTIFICATE**

This is to certify that the thesis entitled “**Implementation of Deep Learning Model for Audio and Video Analysis of Human Emotions**” submitted by Abishek Prakash(19BDS0161),Mallikarjun Hatti(19BCE0889),Thukila. C(19BCT0247), **School of Computer Science and Engineering, VIT**, for the award of the degree of ***Bachelor of Technology in CSE***, is a record of bonafide work carried out by him / her under my supervision during the period, 01.07.2022 to 30.04.2023, as per the VIT code of academic and researchethics.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The thesis fulfills the requirements and regulations of the University and in my opinion meets the necessary standards for submission.

Place: Vellore

Date:



**Signature of the Guide**

**Internal Examiner**

**External Examiner**

**Head of the Department  
Computer Science Engineering**

## ACKNOWLEDGEMENTS

In order to successfully complete our final year project on "**Implementation of Deep Learning Model for Audio and Video Analysis of Human Emotions**", we would like to express our sincere gratitude and appreciation to the people and institutions involved.

First and foremost, we would like to express our gratitude to Dr. Sathya K, our project mentor and advisor, for her unwavering support, inspiration, and advice. Her knowledge and experience have been crucial in forming our opinions and guiding our study.

I would like to express my gratitude to DR.G.VISWANATHAN, Chancellor VELLORE INSTITUTE OF TECHNOLOGY, VELLORE, MR. SANKAR VISWANATHAN, DR.SEKAR VISWANATHAN, MR.GVSELVAM, Vice - Presidents VELLORE INSTITUTE OF TECHNOLOGY, VELLORE, DR.RAMBABU KODALI, Vice-Chancellor, DR.PARTHA SARATHI MALLICK, Pro-Vice Chancellor and Dr. R. Ramesh Babu , Dean, School of Computer Science Engineering(SCOPE), for providing with an environment to work in and for his inspiration during the tenure of the course.

We also want to express our gratitude to our friends and family for their constant support, inspiration, and tolerance throughout our senior project. They gave us the will power to endure the difficult times with their moral support and encouraging remarks.

Last but not least, I express my gratitude and appreciation to all those who have helped me directly or indirectly toward the successful completion of this project and we would like to thank the Almighty for His blessings, protection, and leadership throughout the endeavor.

## **Executive Summary**

Emotions play a crucial role in human communication and have an impact on how we connect with people on a daily basis. Numerous studies have demonstrated the significance of precisely identifying human emotions in a variety of businesses, including gaming, healthcare, and entertainment. Modern models for the automatic detection of human emotions from audio and visual signals have been developed thanks to the recent advancements in deep learning.

In this project, we propose to implement a deep learning model for audio and video analysis of human emotions. A sizable collection of audio and video clips with labeled emotions will be used to train the model. Long Short Term Model (LSTM) and Convolutional Neural Networks (CNNs) are two types of neural networks that can be used to analyze audio and video, respectively. Additionally, in order to optimize previously taught models and decrease training time, we will employ Transfer Learning approaches. The application of this deep learning model for the analysis of human emotions in audio and video gives a potential method for automatic emotion recognition, which has important ramifications for numerous businesses. The concept is well suited for a variety of applications, from gaming and entertainment to healthcare and customer service, thanks to its flexibility and adaptability.

## **TABLE OF CONTENTS**

<b>Serial Number</b>	<b>Topics</b>	<b>Page Number</b>
<b>a)</b>	Acknowledgement	4
<b>b)</b>	Executive Summary	5
<b>c)</b>	Table of Contents	6
<b>d)</b>	List of Figures	7
<b>e)</b>	List of Abbreviations	8
<b>1)</b>	INTRODUCTION	9
<b>1.1)</b>	Theoretical Background	9
<b>1.2)</b>	Motivation	9
<b>1.3)</b>	Aim of the Proposed Work	10
<b>1.4)</b>	Objective of the Proposed Work	11
<b>2)</b>	Literature Survey	13
<b>2.1)</b>	Survey of existing models/work	13
<b>2.2)</b>	Gaps/Summaries Identified	23
<b>3)</b>	Overview of the proposed System	25
<b>3.1)</b>	Introduction and Related Concepts	25
<b>3.2)</b>	Proposed Methodology	25
<b>3.3)</b>	Framework/Architecture/WorkflowModel	32
<b>3.4)</b>	Proposed System Model	32
<b>4)</b>	Proposed System Analysis and Design	36
<b>4.1)</b>	Introduction	41
<b>4.2)</b>	Requirement Analysis	41
<b>5)</b>	Results and Discussion	60
<b>6)</b>	References	67
	APPENDIX	

## List of Figures

Figure Number	Figure Name	Page number
3.1	System architecture for audio emotion recognition	32
3.2	System architecture for video emotion recognition	34
3.3	Proposed System Model for audio emotion recognition	38
3.4	Proposed System Model for video emotion recognition	40
5.1	Average of Human Emotions From Input Video	60
5.2	Graph of Emotions	60
5.3	Output Result for video analysis	61
8)	Output Result for audio analysis	66

## **List of Abbreviations**

Serial Number	Terminology	Abbreviation
1)	CNN	Convolutional Neural Network
2)	KNN	K-Nearest Neighbours
3)	LSTM	Long Short Term Memory
4)	GRU	Gated Recurrent Unit
5)	MLPC	Multi-Layer Perceptron Classifier
6)	FER	Facial Expression Recognizer
7)	RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song



# **1. INTRODUCTION**

## **1.1 THEORETICAL BACKGROUND**

Since their creation, emotion detecting systems have seen substantial development. Early studies in the 1970s and 1980s investigated the use of physiological markers and facial expressions as emotional indicators. The first automated method for identifying emotions from facial expressions was created in the early 1990s. In the late 1990s and early 2000s, machine learning methods including decision trees, support vector machines, and neural networks proliferated, enabling automatic classification of emotions based on information retrieved from physiological data and facial expressions.

Researchers started looking into the use of audio cues like speech and voice tone as additional emotional markers in the middle of the 2000s. In the late 2000s and early 2010s, the integration of several modalities, including facial expressions, physiological signs, and aural signals, enhanced the accuracy of emotion recognition.

In recent years, advanced machine learning algorithms like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been used to detect emotions with high accuracy in real-time applications like virtual assistants, social robots, and video games. Today, as machine learning and computer vision technology expand, emotion detection systems continue to develop, finding use in a variety of industries and playing a bigger role in our digital lives.

## **1.2 MOTIVATION**

The motivation behind our project is to create methods for automatic emotion recognition that are more precise and effective. In disciplines including psychology, neurology, human-computer interaction, and affective computing, among others, emotion recognition is a crucial task.

Emotion recognition has traditionally been done manually by trained specialists using subjective judgments, which can take a lot of time, money, and be biased. Deep learning algorithms have made it feasible to discern emotional states with high accuracy and in real time by automatically analyzing audio and video data.

Numerous real-world uses for automatic emotion identification include marketing research, human-computer interaction, and the diagnosis and treatment of mental illness. Automatic emotion identification, for instance, can be used to track and analyze patients' emotional states throughout therapy sessions, enabling more individualized treatment programs.

Automatic emotion recognition can be used in human-computer interaction to enhance the user experience by customizing the interface to the user's emotional state. For a more engaging experience, a computer game might, for instance, modify the level of difficulty according on the player's emotional state. It can also be utilized in marketing research to examine customer emotional responses to adverts, resulting in more successful advertising campaigns.

### **1.3 AIM OF THE PROPOSED WORK**

The development of precise and effective techniques for automatic emotion recognition is the aim of our proposed work. Deep neural networks are trained to evaluate audio and video data and identify patterns that represent various emotional states. Large datasets, thorough model design and training, and stringent testing and evaluation are all necessary for achieving this goal.

The project's aim is to assess how well various emotion recognition methods function when there is additional noise and shift in the dataset. The goal of the research is to find the most reliable and efficient model for emotion recognition in noisy and shifting data as well as approaches to increase the model's effectiveness.

The project would include choosing and putting different emotion recognition models, including decision trees, KNN, LSTM, GRU, MLPC, and CNN, into practice. To imitate real-world circumstances, the models would be trained and tested on a dataset that has added noise and shift.

Each model's performance would be assessed using a variety of criteria, including accuracy, precision, recall, and F1-score. In order to find the best reliable and efficient model for emotion recognition in noisy and shifting data, the project would compare the performance of various models.

The study would also look into ways to enhance the model's effectiveness by enhancing feature extraction methods, choosing the best hyper parameters, and simplifying the computational complexity of the model. The study would look for methods to raise model productivity without lowering emotion recognition precision.

Overall, the aim of the project is to identify the most effective and efficient model for emotion detection in noisy and shifted data, which would be useful in real-world applications where data is often corrupted or distorted.

## **1.4 OBJECTIVE OF THE PROPOSED WORK**

Our project's goal is to examine the performance of various emotion recognition algorithms in noisy and shifting environments in order to determine which model is the most reliable and accurate under these conditions.

The system would have to be built to withstand the extra noise and data shift in this hypothetical scenario. It would entail choosing and implementing a variety of models, including CNN, KNN, LSTM, GRU, and decision trees, and then training and evaluating them on a dataset with added noise and shift.

By utilizing different models, it will be possible to compare and analyze how well each performs when it comes to identifying emotions in noisy and shifting data. The system would need to include suitable feature extraction methods and noise- and shift-resistant machine learning algorithms.

Under these circumstances, the ultimate goal of an emotion detection system using

several models would be to appropriately categorize emotions despite the additional noise and shift.

In the data. To do this, each model's performance would need to be rigorously tested and validated using a variety of measures, including accuracy, precision, recall, and F1-score.

The goal of this study would be to identify the most reliable and efficient model for emotion recognition in shifting and noisy environments. We can discover the best method for emotion recognition in noisy and shifting environments by evaluating the performance of various models, highlighting the advantages and disadvantages of each model.

The system may also look at ways to enhance the models' effectiveness and efficiency, such as optimizing feature extraction methods, choosing the right hyper parameters, and making the models' computations simpler.

Overall, the objective of an emotion detection system using multiple models in a theoretical aspect where extra noise and shift is added is to identify the most effective and robust model for emotion detection in noisy and shifted data and explore ways to improve the efficiency and effectiveness of the models.

## **2. LITERATURE SURVEY**

### **2.1 SURVEY OF EXISTING MODELS WORK**

- "Limitations of Emotion Recognition from Facial Expressions" (2017) identified challenges in detecting emotions from facial expressions due to variations in the same emotion expressed by different individuals, individual differences in facial anatomy, and the lack of standardized datasets.
- "Automatic Facial Expression Recognition in Standardized and Non-Standardized Emotional Expressions" (2021) proposed a deep learning-based approach to improve facial expression recognition in both standardized and non-standardized emotional expressions, but noted that the performance of the models till depends on the quality of the input images.
- "Deep-Emotion: Facial Expression Recognition using Attentional Convolutional Networks"(2021) presented a facial expression recognition system using attentional CNNs, which outperformed several state-of-the-art methods on standard datasets.
- "Graph based feature extraction and hybrid classification approach for facial expression recognition" (2020) used graph theory for feature extraction and hybrid classification techniques to improve facial expression recognition, achieving better performance than traditional feature extraction methods.
- "Emotion Recognition from facial expression using deep learning" (2019) used deep learning techniques for facial expression recognition and achieved high accuracy on standard datasets, but noted that the models are sensitive to variations in pose and lighting conditions.
- "Speech Emotion Recognition using deep 1D & 2D CNN LSTM networks" (2018) proposed a deep learning-based approach for emotion recognition from speech, achieving high accuracy on standard datasets.
- "Human emotion recognition using deep belief network architecture " (2019) proposed deep learning-based approach for emotion recognition from physiological signals, achieving high accuracy on standard datasets.

- "Development of a Facial Emotion Recognition Method based on combining AAM with DBN"(2015) proposed a facial emotion recognition method combining Active Appearance Models (AAM) with Deep Belief Networks (DBN), achieving high accuracy on standard datasets.
- "Facial emotion recognition using convolutional neural networks (FERC)"(2020) proposed a deep learning-based approach for facial expression recognition, achieving high accuracy on standard datasets.
- "Deep learning-based facial emotion recognition for human–computer Interaction applications"(2021) proposed a deep learning-based approach for facial expression recognition, specifically for human-computer interaction applications.
- "Hybrid-Deep Learning Model for Emotion Recognition Using Facial Expressions" (2020) proposed a hybrid-deep learning model for facial expression recognition, combining the strengths of both deep learning and traditional machine learning techniques, achieving high accuracy on standard datasets.
- "Deep learning approach for emotion recognition from human body movements with feed forward Deep convolution neural networks" (2019) proposed a deep learning-based approach for body movement-based emotion recognition, achieving high accuracy on standard datasets.
- Overall, these works demonstrate the potential of deep learning techniques in emotion detection from different modalities, including facial expressions, speech, and body movements. However, the performance of these models still depends on the quality of the input data and the specific context of emotion detection.

S.No	Title	Year	Algorithms used	Results	Drawbacks
1.	Limitations of Emotion Recognition from Facial Expressions	2017	Decision Tree, Neural Networks and Bayesian Networks	Accuracy – above 90%	<p><b>a)</b> Limited number of participants were present</p> <p><b>b)</b> The metrics and the threshold values used in this project were arbitrarily chosen.</p> <p><b>c)</b> Issues related to the consistency of the emotion recognition channels were not prioritized.</p>
2.	Automatic Facial Expression Recognition in Standardized and Non-Standardized Emotional Expressions	2021	Face Reader, Face++, Azure.	Accuracy – 97%	<p><b>a)</b> There should be sufficient need to do thorough empirical evaluation in order to guide future developments in the computer vision of emotional facial expressions.</p> <p><b>b)</b> The algorithms that were used for the research were not thoroughly examined.</p>
3.	Deep- Emotion: Facial Expression Recognition using Attentional Convolutional Networks.	2021	Regional Attention Network (RAN), Multiple Attention Network (MAN), Deep Self-Attention Network (DSAN)	Overall accuracy is about 92.8%. FER dataset had the highest accuracy of 99.3%	<p><b>a)</b> The datasets that were used in this study started to become challenging.</p> <p><b>b)</b> Suitable framework was not proposed based on the principle of attentional networks.</p>
4)	Graph based feature extraction and hybrid classification approach for facial expression recognition.	2020	Weighted Visibility Graph, GFE-HCA, Neural Networks Classifier	Accuracy- 65%	<p><b>a)</b> Poor recognition of the facial images was done which was not done within accurate landmarks.</p> <p><b>b)</b> There were no significant improvements made in the recognition of facial expressions with a smaller number of image data.</p>

5.	Emotion recognition from facial expression using deep learning	August, 2019	Inception Net – Expression recognition with kaggle datasets.	Accuracy rate - 35.6%( Predicting the emotions). Validation (0.8%) Accuracy achieved.	a) If we do classification and evaluation before conversion we will get more accuracy rate.  b) Less accuracy rate (Transfer learning-train the model, generated less no. of spectrogram for training, to Less accuracy. Less dataset for training process.)
6.	Speech Emotion Recognition using deep 1D& 2D CNN LSTM networks	August 2018	2DCNNLSTM Accuracy-95.33% and 95.89% (Berlin EmoDB-->speaker dependent ,speaker independent) IEMOCAPDB -->speaker-dependent & independent)	Speaker dependent 89.16 Independent 52.14  DBN and CNN - 73.78%, 40.02%	a).the designed networks recognize the emotion cannot be explained in more detail meaning the “black box” of these networks have not been uncovered.  b)new optimization algorithms
7.	Deep learning approach for emotion recognition from human body movements with feed forward Deep convolution neural networks	2019	GEMEP dataset(audio and video recording) ,CNN	2015- Gavrilesu 86.4% 2016 – Radoslaw 73.0% 2017 – Nourhan, 90.7%	This work 95.4%
8.	Human emotion recognition using deep belief network architecture	2019	Gaussian Support Vector Machine,(FGSV M) DEAP dataset.	Happy –100%  Accuracy -89.53%	a)better than the radio frequency based emotion analyzer with 72%, the SVM based emotion classifier 82.9%, Naïve Bayes classifier based DEAP 65.1% and 61.8%



9	Development of a Facial Emotion Recognition Method based on combining AAM with DBN	2015	Combining AAM Facial Action Coding System Facial feature extraction method based on combining AAM with FACS.	Average Accuracy – 52%	<p>It will be the subject of future work to apply advanced feature extraction methods because of the vulnerability of an AAM such as the necessity for many landmarks and time data.</p> <p>Also, in order to recognize emotion states more delicately, it is required that the advanced structure, and parameter learning of the Bayesian Network.</p>
10	Facial emotion recognition using convolutional neural networks(FERC)	2020	The FERC algorithm for face detection in a single CNN network	Accuracy – 92%	<p>we only have considered five moods to classify, the sixth and seventh mood cases were misclassified, adding to the error</p> <p>Achieved maximum accuracy up to 99.3% but at the cost of 22 layers of neural network.</p>
11.	Deep learning-based facial emotion recognition for human-computer interaction applications	2021	VGG19 model, Mobile Netv2	Accuracy- 94.2%	<p>Among all four pre-trained networks, MobileNet achieved the highest accuracy.</p> <p>The accuracy achieved using the VGG19 model is 96%, Resnet50 is 97.7%, Inception V3 is 98.5%, and MobileNet is 94.2%</p>
12.	Hybrid-Deep Learning Model for Emotion Recognition Using Facial Expressions	2020	FER2013 and JAFFE datasets	97.07% and 94.12%	<p>The model cannot extend to classify the primary and secondary emotions on real-time video data and images.</p>

13	Speech Emotion Recognition Using Deep Learning Techniques	2019	CNN , KNN, GMM, HMM and SVM classifier.	Accuracy- 92%	If the data's were varying Then the efficiency can below.
14	Multi modal Emotion Recognition using Deep Learning	2021	LSTM, CNN and SVM	LSTM- 73%,SVM- 72%	Efficiency of the results will be high when we using only one algorithm.  In multi model , results will below if we propose a new approaches or new algorithm to the datasets already trained
15	Using Deep Convolutional Neural Network for Emotion Detection on a Physiological Signals Dataset	2018	AMIGOS Dataset ,DCNN ,Naïve Bayes	58%and71%	More number of instances gives more accuracy and efficient.  Less number of instances of a dataset gives low accuracy.
16	Speech Emotion Recognition using Convolutional and Recurrent Neural Networks	2018	STFT,CNN,LSTM	Accuracy - 85%	It's hard to insert the text features into SVM.  When the Emotion labels not stable then SER system won't work properly. It leads to give accuracy low.

17	Audio and face video emotion recognition in the wild using deep neural networks and small datasets	2016	CNN, LSTM , Transfer Learning	Accuracy – 53.9%	<p>When the dataset is small it works efficiently.</p> <p>It won't give better accuracy when large datasets trained.</p> <p>It depends on the dataset what we used.</p>
18	Deep Learning-Based Emotion Recognition from Real-Time Videos	2019	CNN,DNN-VGGs Net and Caffe	Overall Accuracy - 60%	<p>The webcam resolution should be 1920 X 1080.</p> <p>It's failed to detect the expected emotion.</p> <p>Didn't find the way to give the desired emotional state</p>
19	Face Recognition from Video using Deep Learning	2020	CNN, FaceNet	90%	<p>This model can recognize the side face and blur red image also</p> <p>,still it can give the better accuracy.</p>
20	Facial Expression Recognition via Deep Learning	2017	RAFD, CAFFE,CK+	Accuracy – 71.4%	<p>We have used CAFFE – UBUNTU14.04version,GT-GPU- 2GB memory .We can't use the multiple hidden layers for emotion recognition.</p>

## **2.2 GAPS/SUMMARIES IDENTIFIED**

The papers address the issues of emotion recognition using various approaches and techniques. However, they also have certain limitations.

- "Limitations of Emotion Recognition from Facial Expressions 2017" identifies that existing facial expression recognition systems have limitations in accounting for cultural and individual differences in facial expressions. The paper suggests the need for more comprehensive datasets that can capture these differences to improve the performance and generalization of the models.
- "Automatic Facial Expression Recognition in Standardized and Non-Standardized Emotional Expressions 2021" presents a method for automatic facial expression recognition. However, the paper does not address the issue of cultural and individual differences in facial expressions, which can affect the accuracy and reliability of the models.
- "Deep- Emotion: Facial Expression Recognition using Attentional Convolutional Networks 2021" focuses on using attentional convolutional networks to improve the accuracy of facial expression recognition. However, the paper does not address the issue of individual and cultural differences in facial expressions, which can affect the generalization of the models.
- "Graph based feature extraction and hybrid classification approach for facial expression recognition 2020" proposes a graph-based feature extraction method and a hybrid classification approach to improve the accuracy of facial expression recognition. However, the paper does not address the issue of individual and cultural differences in facial expressions, which can affect the generalization of the models.
- "Speech Emotion Recognition using deep 1D & 2D CNN LSTM networks August 2018" proposes a deep learning approach for speech emotion recognition using 1D and 2D CNN LSTM networks. However, the paper does not address the issue of variability in speech patterns due to cultural and individual differences.

- "Deep learning approach for emotion recognition from human body movements with feed forward Deep convolution neural networks 2019" proposes a deep learning approach for emotion recognition from human body movements using feed forward deep convolution neural networks. However, the paper does not address the issue of individual and cultural differences in body movements.
- "Human emotion recognition using deep belief network architecture 2019" focuses on using deep belief network architecture for emotion recognition from facial expressions. However, the paper does not address the issue of individual and cultural differences in facial expressions.
- "Development of a Facial Emotion Recognition Method based on combining AAM with DBN 2015" proposes a method for facial emotion recognition that combines Active Appearance Models (AAM) with DBN. However, the paper does not address the issue of individual and cultural differences in facial expressions.
- "Facial emotion recognition using convolutional neural networks (FERC) 2020" focuses on using convolutional neural networks for facial emotion recognition. However, the paper does not address the issue of individual and cultural differences in facial expressions.
- "Deep learning-based facial emotion recognition for human–computer interaction applications 2021" proposes a deep learning-based facial emotion recognition approach for human-computer interaction applications. However, the paper does not address the issue of individual and cultural differences in facial expressions.
- "Hybrid-Deep Learning Model for Emotion Recognition Using Facial Expressions 2020" proposes a hybrid deep learning model for facial expression recognition using both static and dynamic features. However, the paper does not address the issue of individual and cultural differences in facial expressions.
- In summary, while the papers present various approaches and techniques for emotion recognition from different modalities, they do not address the issue of individual and cultural differences in expressions, which can affect the accuracy, reliability, and generalization of the models. Therefore, there is a need for more comprehensive datasets that can capture these differences and more robust and accurate models that can handle them.

### 3. OVERVIEW OF THE PROPOSED SYSTEM

#### 3.1 INTRODUCTION

From the Literature paper we have concluded the drawbacks and with the help of that we are proposing a new model in our paper. In this paper, we propose a new architecture based on Convolutional Neural Network(CNN) and Long Short Term Memory(LSTM) for the audio-visual recognition that classifies six basic expressions with neutral, happiness, anger, disgust, fear, sadness, and surprise. In addition, we used well-trained models like GRU and MLPC classifier to recognize patterns in extracted facial image features like the shape and position of the brows, mouth, and eyes, and assign weights to these patterns. This integrated approach is divided into two stages. In the first stage, CNN extracts visual features first, and then LSTM is used to bind the relationship between image sequences and emotions. In the second stage, GRU processes sequential data which can be used to recognize emotions using audio and video signals. It processes the input vectors and generates an output corresponding to the person's emotional state. MLPC classifier analyzes the patterns and assigns a probability score to each emotion category. The predicted emotion is the emotion with the highest probability. Furthermore, the outcome of this architecture is evaluated using confusion matrices and compared to relevant architectures and well-known FER datasets.

These models analyze audio and video data to extract relevant features like pitch, tone, intensity, facial expressions, and body language. These models are then trained to recognize specific emotions using these features.

#### 3.2 PROPOSED METHODOLOGY:-

##### **For audio:**

Certainly! Here are some equations that are commonly used in the methodology described:

1. Data Preprocessing:

**-LSTM for sequential data preprocessing:**

-Input sequence:  $X=[x_1, x_2, \dots, x_n]$

- Hidden state at time step t:

i)  $h_t = \text{LSTM}(x_t, h_{t-1})$

- Output at time step t:

ii)  $o_t = \text{Activation}(W_o h_t + b_o)$

## 2. Data Augmentation:

### **-Adding white noise:**

#### **-Augmented waveform:**

iii)  $x_{\text{augmented}} = x + \alpha * \varepsilon$

- where exist the original waveform,  $\alpha$  is a noise factor, and  $\varepsilon$  is white noise sampled from a Gaussian distribution.

### **- Pitch shifting:**

#### **- Augmented waveform:**

iv)  $x_{\text{augmented}} = \text{Stretch}(x, \alpha)$

- where  $\alpha$  is a pitch shifting factor.

### **- Time shifting:**

#### **- Augmented waveform:**

v)  $x_{\text{augmented}} = \text{Shift}(x, \alpha)$

- where  $\alpha$  is a time shifting factor.

### **- Speed perturbation:**

#### **- Augmented waveform:**

vi)  $x_{\text{augmented}} = \text{Resample}(x, \alpha)$

- where  $\alpha$  is a speed perturbation factor.

## 3. Feature Extraction:

### **- CNN:**

#### **- Convolutional layer output:**

vii)  $y = \text{Convolution}(x, w) + b$

- where exist the input spectrogram or MFCC image,  $w$  is the convolutional filter, and  $b$  is the bias term.

### **- MLP:**

#### **- Forwardpass:**

-  $y = \text{Activation}(W_1 x + b_1)$

- where exist the flattened spectrogram or MFCC image,  $W_1$  is the weight matrix, and  $b_1$  is the bias term.

#### 4. Model Training:

##### - **Decision Tree:**

- Splitting criterion : G in impurity or entropy
- Decision rule: if  $feature_i < threshold$ , go to the left child; otherwise, go to the right child.

##### - **KNN:**

- Distance metric: Euclidean distance or other distance measures.
- Voting scheme: Majority voting or weighted voting.

##### - **MLP Classifier:**

- Forward pass:  
viii)  $y = \text{Activation}(W_1x + b_1)$
- where exist the input feature vector,  $W_1$  is the weight matrix, and  $b_1$  is the bias term.

##### - **GRU:**

- Update gate:  
ix)  $z_t = \sigma(W_u u_t + b_u + W_v h_{t-1} + b_v)$
- Reset gate:  
x)  $r_t = \sigma(W_r u_t + b_r + W_v h_{t-1} + b_v)$
- New memory content:  
xi)  $h_t = \tanh(W_u u_t + b_u + r_t * (W_v h_{t-1} + b_v))$
- Hidden state at time step t:  
xiii)  $h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$

##### - **LSTM:**

- Input gate:  
xiv)  $i_t = \sigma(W_i u_t + b_i + W_i h_{t-1} + b_i + W_i c_{t-1} + b_i)$
- Forget gate:  
xv)  $f_t = \sigma(W_h u_t + b_h + W_h h_{t-1} + b_h + W_h c_{t-1} + b_h)$



➤ Cell state:

$$\text{xvi) } c_t = f_t * c_{t-1} + i_t$$

**For video:**

The proposed methodology for visual emotion detection system is as follows:

1. Forward Pass Equation:

In the forward pass, the output of each layer in the CNN can be calculated as follows:

$$\text{xvii) } Z = W * X + b$$

$$\text{xviii) } A = \text{activation}(Z)$$

Where:

- Z is the weighted sum of inputs,
- W is the weight matrix of the layer,
- X is the input to the layer,
- B is the bias vector of the layer,
- activation() is the activation function applied element-wise to Z,
- A is the output or activation of the layer.

2. Loss Function:

The loss function measures the discrepancy between the predicted emotion and the true emotion label. A common choice is the categorical cross-entropy loss:

$$\text{xix) } \text{Loss} = -\sum (y_{\text{true}} * \log(y_{\text{pred}}))$$

Where:

- $y_{\text{true}}$  is the true emotion label (one-hot encoded),
- $y_{\text{pred}}$  is the predicted probability distribution over the emotion classes.

3. Back propagation:

During the training phase, the gradients of the loss function with respect to the parameters (weights and biases) of the model are computed using back propagation. The gradients are then used to update the parameters to minimize the loss. The back propagation equations for a single layer are as follows:

$$\text{xx) } dZ = A - y_{\text{true}}$$

$$\text{xxi) } dW = (1/m) * dZ * X.T$$

$$\text{xxii) } db = (1/m) * \text{np.sum}(dZ, \text{axis}=1, \text{keepdims}=\text{True})$$

$$\text{xxiii) } dX = W.T * dZ$$

Where:

- $dZ$  is the gradient of the loss with respect to  $Z$ ,
- $dW$  is the gradient of the loss with respect to the weights,
- $db$  is the gradient of the loss with respect to the biases,
- $dX$  is the gradient of the loss with respect to the inputs,
- $m$  is the number of training examples.

These equations are applied iteratively, starting from the output layer and propagating backward through the layers of the CNN.

The above equations represent a general overview, and the specific architecture and variations of the CNN may lead to additional or modified equations. The choice of activation function, optimization algorithm, and other factors can also influence the equations used in the methodology.

### 3.3 Framework, Architecture or Module for the Proposed System(with explanation)

**For audio:**

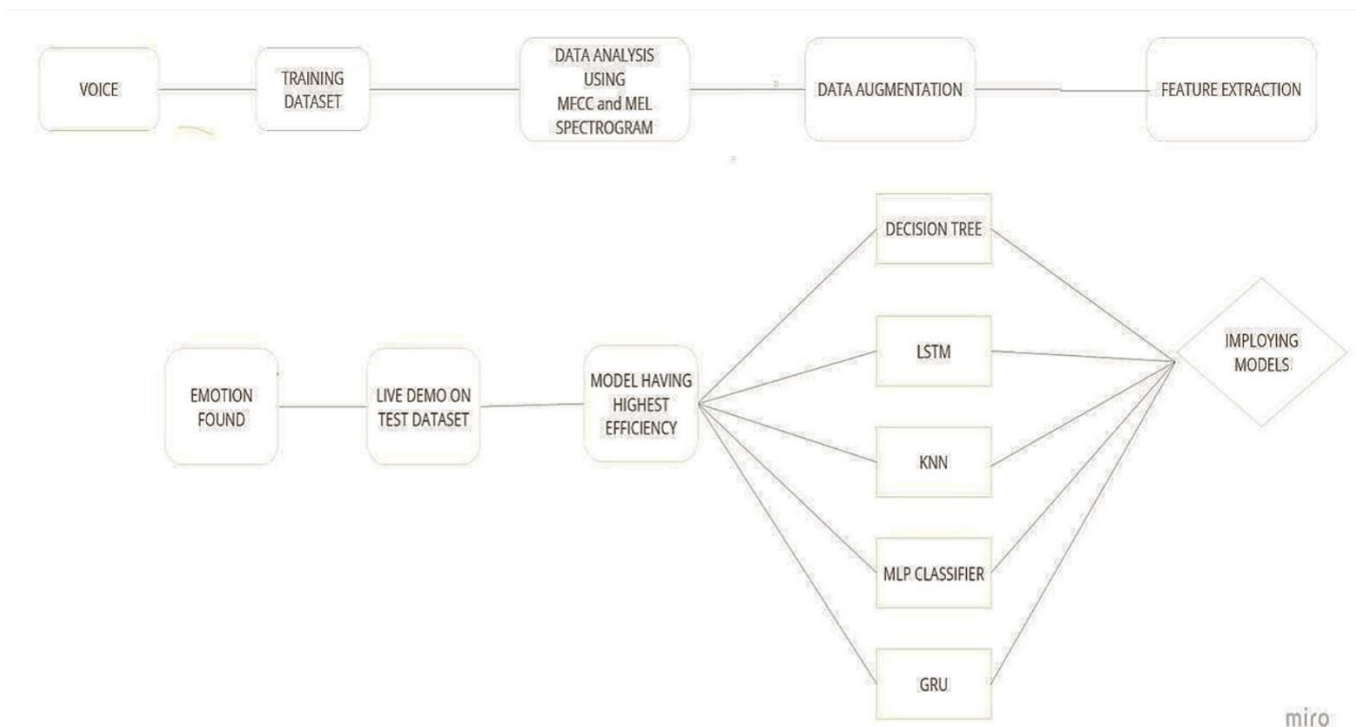


Figure 3.1: System architecture for audio emotion recognition

### 1. Data Preparation:

- Import the necessary libraries for audio analysis and feature extraction.
- Set the path to the RAVDESS dataset directory.
- Define emotions in the dataset and print their count.

### 2. Exploratory Data Analysis:

- Plot a bar graph to visualize the distribution of emotions in the dataset.
- Create a log mel spectrogram for specific emotions (e.g., 'Female Happy', 'Female Sad', 'Male Happy', 'Male Sad') to analyze the localized frequency content of the audio signals in the mel frequency scale.

### 3. Data Augmentation:

- Apply data augmentation techniques to create additional training samples.
- Use techniques like noise injection and time shifting to introduce small disturbances and generate augmented data.

### 4. Feature Extraction:

- Define a function to extract features from the audio data.
- Extract key features such as MFCC (Mel Frequency Cepstral Coefficients) and Mel Spectrogram from the audio files.
- Save the extracted features and corresponding labels in a CSV file (e.g., 'emotion.csv').

### 5. Data Processing:

- Split the dataset into training and test sets.
- Use LSTM (Long Short-Term Memory) for training and testing the model.

### 6. Model Selection and Evaluation:

- Employ various models such as Decision Tree, KNN (K-Nearest Neighbors), MLP Classifier (Multi-Layer Perceptron), GRU (Gated Recurrent Unit), LSTM, and CNN (Convolutional Neural Network).
- Evaluate the models' efficiency and performance metrics.
- Select the most efficient model based on the evaluation results (e.g., CNN with a accuracy of 75.55%).

-

### 7. Model Training and Saving:

- Save the trained model with the name 'modelh3.h5' for future use.
- Load the trained model weights into a new model.

### 8. Live Emotion Recognition:

- Load the input audio file.
- Print the frequency graph of the audio signal to visualize its waveform.
- Resample the audio to the desired sample rate.
- Extract MFCC features from the resampled audio.
- Preprocess the features and prepare them for speech emotion recognition.
- Make predictions on the input data using the trained model.
- Print the predicted emotion label.

The system architecture involves steps such as data preparation, exploratory data analysis, data augmentation, feature extraction, data processing, model selection and evaluation, model training and saving, and live emotion recognition. Each step plays a crucial role in building an effective emotion recognition system using speech.

### For video:

#### SYSTEM DIAGRAM:

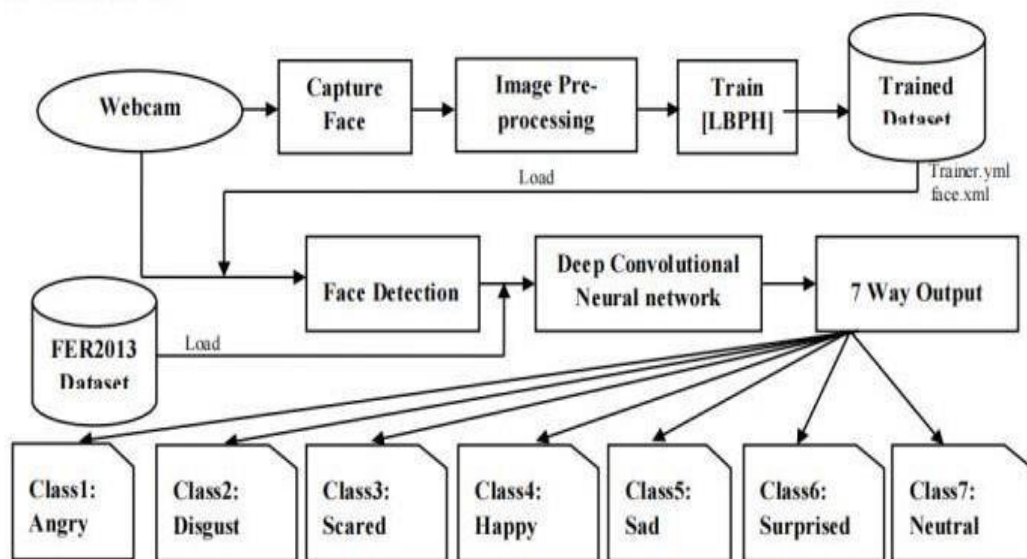


Figure 3.2: System architecture for video emotion recognition

#### 1. Video and Capture Face:

- The system begins by accessing the video to capture live video frames.
- The Capture Face module captures the face from each frame using a face detection algorithm. This module detects and extracts the face region of interest (ROI) from the webcam feed.

#### 2. Image Preprocessing:

- The captured face images are preprocessed to enhance the quality and remove any noise or inconsistencies.
- Preprocessing techniques may include resizing the images to a fixed size, normalizing pixel values, and applying filters for noise reduction.

#### 3. Train Data:

- A training dataset is prepared to train the emotion recognition model.
- The dataset consists of pre-labeled face images, where each image is associated with a specific emotion label.
- The training data is divided into input images (captured faces) and their corresponding emotion labels.

#### 4. Face Detection:

- The preprocessed face images from the webcam are fed into a face detection module.
- This module utilizes a face detection algorithm to identify and locate faces within the images.
- The detected faces are then passed to the emotion recognition module for further analysis.

#### 5. CNN (Convolutional Neural Network):

- The emotion recognition module employs a Convolutional Neural Network (CNN) model to classify the emotions present in the detected faces.
- The CNN model is trained using the training dataset, where it learns to extract meaningful features from the face images and predict the corresponding emotion labels.
- The model architecture typically consists of multiple convolutional layers, pooling layers, and fully connected layers.

#### 6. 7 Output Emotions:

- The CNN model is designed to predict one of the seven predefined emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral.
- The model analyzes the features extracted from the detected face and assigns a probability distribution over the seven emotions.

## 7. Final Output:

- The system displays the predicted emotion label based on the highest probability from the CNN model.
- The output can be presented in real-time, associating the predicted emotion label with the captured face on the video feed.

This system architecture showcases the process of emotion recognition using a video feed. It involves capturing faces from the webcam, preprocessing the images, training a CNN model on labeled data, detecting faces in the captured images, and predicting emotions using the CNN model. The architecture focuses on real-time emotion recognition from live video input.

### 3.4. Proposed System Model

**(ERDiagram /UMLDiagram/Mathematical Modeling)**

#### **For audio:**

The proposed system model for the above methodology is as follows:

```
import librosa
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
dataset_path = '/path/to/dataset'
def print_emotions_count():
    emotions = ['Neutral', 'Happy', 'Sad', 'Angry', 'Fearful', 'Disgust', 'Surprised']
    counts = [count_emotions(emotion) for emotion in emotions]
    for emotion, count in zip(emotions, counts):
        print(f"{emotion}: {count}")
def plot_emotions_count():
    emotions = ['Neutral', 'Happy', 'Sad', 'Angry', 'Fearful', 'Disgust', 'Surprised']
    counts = [count_emotions(emotion) for emotion in emotions]
    plt.figure(figsize=(10, 6))
    sns.barplot(x=emotions, y=counts)
    plt.xlabel('Emotions')
    plt.ylabel('Count')
    plt.title('Emotions Distribution')
    plt.show()
def create_log_mel_spectrogram(emotions):
    for emotion in emotions:
```

```

    audio_path = f'{dataset_path}/{emotion}_audio.wav'
    audio, sr = librosa.load(audio_path, sr=None)
    mel_spec = librosa.feature.melspectrogram(audio, sr=sr, n_mels=128, hop_length=512)
    log_mel_spec = librosa.amplitude_to_db(mel_spec)
def extract_features(audio):
    # Extract MFCC features
    mfcc = librosa.feature.mfcc(audio, sr=sr, n_mfcc=13)
    delta_mfcc = librosa.feature.delta(mfcc)
    delta2_mfcc = librosa.feature.delta(mfcc, order=2)
    features = np.concatenate((mfcc, delta_mfcc, delta2_mfcc), axis=0)
    return features
#Performing data augmentation techniques such as noise injection and time shifting to
#Create augmented data samples.
#Extracting features from audio data:
# Saving features to 'emotion.csv'
def save_features(features, labels):
    data = pd.DataFrame(features)
    data['label'] = labels
    data.to_csv('emotion.csv', index=False)
#Preprocessing the data by splitting it into training and test sets and using LSTM for
training and testing.
#Evaluating various models (Decision Tree, KNN, MLP Classifier, GRU, LSTM,
CNN) and selecting the most efficient one (e.g., CNN with 75.55% accuracy).
#Saving the trained model
# Load the trained model
model = keras.models.load_model('modelh3.h5')

# Load the input audio file
audio_path = '/path/to/input/audio.wav'
audio, sr = librosa.load(audio_path, sr=None)

# Print the frequency graph of the audio
plt.figure(figsize=(10, 4))
librosa.display.waveplot(audio, sr=sr)
plt.title('Input Audio')
plt.xlabel('Time')
plt.ylabel('Amplitude')
plt.show()

```

```

# Resample audio to the desired sample rate
resampled_audio = librosa
desired_sr = 22050
resampled_audio = librosa.resample(audio, sr, desired_sr)
mfcc_features = librosa.feature.mfcc(resampled_audio, sr=desired_sr, n_mfcc=13)
# Perform feature scaling
scaled_features = (mfcc_features - np.mean(mfcc_features)) / np.std(mfcc_features)
# Reshape the features array to match the expected input shape of the model
input_features = scaled_features.reshape(1,scaled_features.shape[0]
, scaled_features.shape[1], 1)
#Making predictions using the trained model:
predictions = model.predict(input_features)
emotion_labels = ['Neutral', 'Happy', 'Sad', 'Angry', 'Fearful', 'Disgust', 'Surprised']
predicted_emotion = emotion_labels[np.argmax(predictions)]
# Print the predicted emotion
print("Predicted Emotion:", predicted_emotion)

```

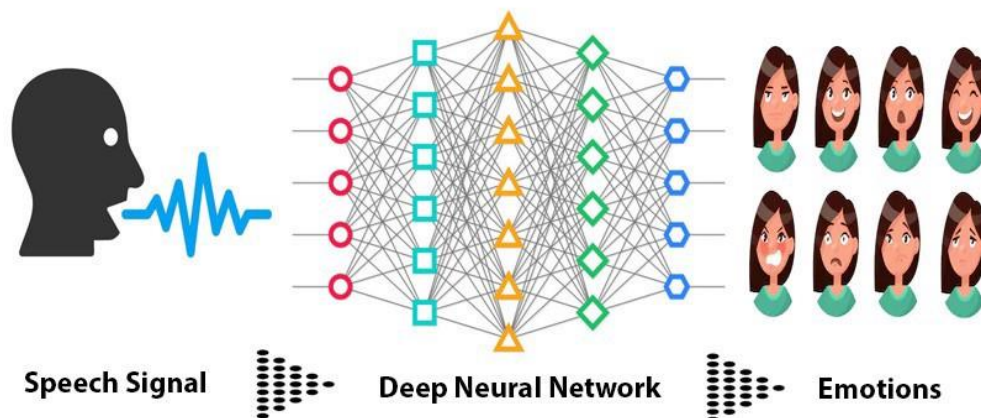


Figure 3.3:Proposed system model for audio analysis

#### **For video:**

The proposed system model for emotion detection system for video using CNN:

Step 1: Load and Preprocess the Video Dataset

```
[X_train, y_train, X_test, y_test] = LoadVideoDataset()
```

Step 2: Data Preprocessing

```
X_train = PreprocessFrames(X_train)
```

```
X_test = PreprocessFrames(X_test)
```

Step 3: Build CNN Model Architecture

```
model = BuildCNNModel()
```



#### Step 4: Train the CNN Model

```
max_epochs = 100
current_epoch = 0
while current_epoch < max_epochs:
    TrainModel(model, X_train, y_train)
    current_epoch += 1
```

#### Step 5: Evaluate the Model

```
accuracy_threshold = 0.9
accuracy = EvaluateModel(model, X_test, y_test)
while accuracy < accuracy_threshold:
    TrainModel(model, X_train, y_train)
    accuracy = EvaluateModel(model, X_test, y_test)
```

#### Step 6: Fine-tune and Optimize the Model

```
max_iterations = 10
current_iteration = 0
while current_iteration < max_iterations:
    model = FineTuneModel(model, X_train, y_train)
    current_iteration += 1
```

#### Step 7: Emotion Prediction for New Videos

```
new_video_frames = LoadNewVideoFrames()
preprocessed_frames = PreprocessFrames(new_video_frames)
predicted_emotions = model.predict(preprocessed_frames)
```

#### Step 8: Post-processing and Visualization

```
smoothed_emotions = SmoothEmotions(predicted_emotions)
VisualizeEmotions(new_video_frames, smoothed_emotions)
```

#### Step 9: Deploy and Use the Emotion Detection System

```
DeployEmotionDetectionSystem(model)
```



Figure 3.4: Proposed system model for video analysis

## **4.PROPOSED SYSTEM ANALYSIS AND DESIGN**

### **4.1 INTRODUCTION**

This system aims to help people recognize and analyze emotions, which can be useful in various fields, including healthcare, psychology and marketing.

The system's design would involve collecting data on human emotions through various sources, including physiological sensors, facial recognition software, or self-reporting surveys. The data would then be curated and labeled to train machine learning algorithms to recognize different emotions accurately.

The system would also involve feature extraction techniques to identify patterns in the data, such as changes in facial expression, voice tone, or physiological responses. These features would be used as input for the machine learning algorithms, which would classify and interpret the emotional states of individuals accurately.

The system's output would be presented through an intuitive and user-friendly interface, allowing users to access and analyze the emotional data easily. This system could have potential applications in various industries, including healthcare, where it could help diagnose mental health disorders, and marketing, where it could be used to understand consumers' emotional responses to products or advertisements.

### **4.2 REQUIREMENT ANALYSIS**

#### **4.2.1 FUNCTIONAL REQUIREMENTS**

##### **4.2.1.1 PRODUCT PERSPECTIVE**

An emotion detection system that uses audio and video analysis can offer a number of advantages and potential applications from a products and point. Here are some considerations:

1) Target Audience: Researchers investigating emotions, marketers seeking to gauge consumer sentiment, mental health practitioners, researchers interested in emotions, and those developing digital assistants or robots that interact with people could all benefit from emotion detection systems.

- 2) Technical prerequisites: To effectively recognize and categorize emotions, the system would need to use cutting-edge audio and video analysis methods, such as machine learning algorithms, computer vision, and natural language processing.
- 3) User Interface: The system's user interface may be created to be straight forward and user-friendly depending on the intended audience, or it may be more intricate to enable in-depth emotion research.
- 4) Data security and privacy: Privacy and security will be major issues with any system that incorporates audio and video data, as they always are. To protect user data, the product would need to adhere to all applicable data protection laws and have strong security mechanisms in place.
- 5) Arrangement with other systems: To deliver useful insights, the system might need to link with other programs, such as electronic health records or socialmedia analytics tools, depending on the audience it is intended for.
- 6) Cost: The price will vary depending on the system's complexity, the size of the development team, and the need for continuing maintenance and support.

#### **4.2.1.2 PRODUCT FEATURES**

1) Real-time emotion analysis and detection:

The system would be capable of identifying and analyzing emotions in real-time, giving users immediate feedback.

2) Multi modal analysis:

To provide a more precise and thorough analysis of emotional state, the system would use both audio and video analysis.

3) Advanced machine learning algorithms

This would be used by the system to precisely distinguish and categorize a variety of emotions, such as happiness, sadness, rage, surprise, and more.

4) User-customizable settings:

Users can choose the emotions they want the system to recognize and change the sensitivity of the emotion detection.

5) Integration with current platforms:

The system might be quickly integrated with current collaboration and communication tools, such social networking or video conferencing applications.

6) User feedback:

Depending on the user's emotional state, the system could offer suggestions for how to improve communication.

7) Privacy and data security:

The system would be created with these concepts in mind, and it would have strong data protection measures in place to guarantee that user data is safe and secure.

8) Analytics and reporting:

The system might offer tools for tracking and analyzing emotional patterns and trends overtime for enterprises.

### **4.2.1.3 USER CHARACTERISTICS**

Users need to examine emotional expressions in audio and video data who need accurate interpretation and real-time analysis of emotional cues.

1. Clinicians:

Deep learning models such as CNN,LSTM and RNN can be used by clinicians ,including psychiatrists and psychologists, to analyze audio and video of human emotions in order to diagnose and treat mental health issues like depression, anxiety, and bipolar disorder.

## 2. Entrepreneurs in the entertainment industry:

In order to produce more engrossing and immersive material, members of the entertainment industry, such as filmmakers and game developers, may use audio and video analysis of human emotions.

## 3. Customer Service Personnel:

Customer service agents may use our project to better comprehend client feedback and address the requirements of the customers.

### **4.2.1.4 ASSUMPTION AND DEPENDENCIES**

#### Assumptions:

- The availability of huge and diverse datasets of audio and video recordings that faithfully capture human emotions is necessary for the deep learning models to function effectively.
- The underlying premise of deep learning models is that an individual's emotional responses remain constant across settings and circumstances.
- The models also presuppose that all cultures and races have the same facial and vocal expressions of human emotions.
- The models presuppose that discrete categories like joyful, sad, furious, etc. can accurately categorize emotions.

#### Dependencies:

- The efficiency of deep learning models for analyzing human emotions in audio and video depends on the computing power, memory, and storage that are available.
- These models' performance and accuracy are also influenced by the accuracy and resolution of the input data, which includes the lighting, camera angle, and recording quality.
- The particular environment and application situation in which deep learning models are used may have an impact on their generalization and efficacy.
- The subjective interpretation of human emotions by the labelers or annotators of the training data may potentially have an effect on the models' performance.

## 4.2.1.5 DOMAIN REQUIREMENTS

### 1. Data Gathering:

It takes a lot of different types of data with tagged emotions to train a deep learning model for emotion analysis. The information should document the many emotions that are shown through speech, body language, and facial expressions.

### 2. Signal Processing:

Preprocessing is required to turn the raw data into a useful representation that deep learning models can use. While video data needs to be divided and preprocessed before input into the model, audio data needs methods like Fourier transformation and MFCC for feature extraction.

### 3. Machine learning algorithms:

For deep learning model training, familiarity with a variety of machine learning techniques is necessary. Knowledge of supervised, unsupervised, and reinforcement learning methods falls under this category.

### 4. Neural networks:

A solid knowledge of neural network architecture, including recurrent neural networks(RNNs) for speech recognition and natural language processing and convolutional neural networks (CNNs)for image processing.

### 5. Programming language skills:

Understanding of programming languages like Python, Matlab, R, and libraries like TensorFlow, PyTorch, and Keras is necessary for implementation.

### 6. Psychology:

The results and the model's assessment both require a grasp of psychology and human behavior.

## **4.2.1.6 USER REQUIREMENTS**

### **1. Precision and Dependability:**

High accuracy and reliability are required in the identification and interpretation of emotional expressions by users of our project. The models must be capable of handling a range of emotional emotions in different situations and civilizations.

### **2. Real-time Performance:**

Users need deep learning models to deliver real-time performance with little latency in some applications, such real-time emotion identification in video conferencing or virtual reality environments.

### **3. User Interface:**

Users need an interface that is simple to use and intuitive so they can readily engage with deep learning models. On the identification and interpretation of emotional expressions, the interface should offer brief, clear feedback.

### **4. Customizability:**

Users might need the ability to alter deep learning models to meet their unique needs. This can entail altering the model's design, using their own data to train the model, or integrating the model with their current systems.

### **5. Scalability:**

Users might need deep learning models to be scalable in order to handle enormous data volumes or high usage rates. Distributed computing or cloud-based services might be needed for this.

### **6. Security and Confidentiality:**

Users demand reassurance that the deep learning models will maintain the privacy and security of their data. This can necessitate the use of encryption, safe data storage, and compliance with pertinent privacy laws.



#### 7. Efficiency:

The models should be resource-efficient, work on computers with modest processing power, and work on mobile devices like smartphones, tablets, and wearable technology.

#### 8. Versatility:

The models ought to be flexible enough to operate in a variety of audio and video formats, languages, and cultural situations.

#### 9. User-friendly:

The output of the models should be simple to understand, and the user interface should be simple for the target users to operate.

### **4.2.2 NON-FUNCTIONAL REQUIREMENTS**

Non-functional requirements determine the resources required, time interval, transaction rates, throughput, and everything that deals with the performance of the system.

#### **4.2.2.1 PRODUCT REQUIREMENTS**

##### **4.2.2.1.1 Maintainability**

The thesis put out by our team can have a number of effects on maintainability, which refers to how quickly and easily a system can be updated, fixed, or modified over the course of its existence. Here are some factors:

1) Quality of data: The effectiveness and maintainability of the emotion detection system can be significantly impacted by the quality and quantity of data used for training and testing. It is crucial to gather and curate high-quality data that is representative of the target audience and emotion states in order to ensure maintainability. This may lessen the need for frequent system modifications or retraining.

2) Feature extraction: When utilizing audio and video analysis to detect emotions, feature extraction is a crucial step. The system's performance may be impacted by the caliber of features that are derived from audio and video data. As a result, it's crucial to pick reliable feature extraction methods that are pertinent to the objective of emotion identification.

3) Selection of algorithm: The system's ability to be maintained can be impacted by the algorithm that is used to identify emotions. While some algorithms are simpler and easier to apply, others are more complex and demand more resources and specialized knowledge to adapt or update.

4) Scalability: The emotion detection system needs to be scalable in order to accommodate growing data and traffic volumes over time. This can ensure that the system maintains its effectiveness and efficiency even if the amount of data or users grows.

5) Version control and documentation: Adequate version control and documentation can help guarantee system maintainability over the system's lifespan. Version control can track changes and encourage developer cooperation, while clear documentation can make it simpler for developers to comprehend the system's architecture and implementation.

#### **4.2.2.1.2 Portability**

With the help of our proposed thesis, the ease and effectiveness with which a system may be relocated or modified to fit other hardware or software settings can have an impact on portability, which is a term used to describe the capacity to identify emotions using audio-visual signals. Here are some portability features that are quite helpful to our project:

1) Hardware needs: Depending on the algorithm's complexity and the volume of data being analyzed, the hardware requirements for emotion identification utilizing audio and video analysis can change. It's critical to choose hardware that is widely accessible and compatible with a range of software environments.

2) Software requirements: The emotion detection system's software requirements may limit its portability. It should be made as easy as possible to move the system to multiple contexts by reducing the system's reliance on certain software libraries or frameworks.

3) Complexity of the algorithm: Complex algorithms maybe less portable than simple algorithms because they may need specialized hardware or software environments to function effectively.

4) Data format: The format of the input data used for emotion detection has a great impact on the portability of the system. To ensure interoperability with various surroundings, the system should be designed to accommodate a number of input data types, including well

known audio and video file formats.

- 5) Infrastructure for deployment: Lastly, the infrastructure utilized to implement the emotion detection system may be directly proportional to the portability of the system. The system should be built with the flexibility to handle diverse hardware and software configurations and be simple to install in various contexts.

#### **4.2.2.1.3 Performance**

Performance, which is the speed, accuracy, and effectiveness with which a system can carryout its intended purpose, can be profoundly affected with the help of our proposed system. These include the following factors that affect performance:

- 1) Accuracy of the algorithm: A product's fulfillment can be greatly affected by the algorithm's accuracy in detecting emotions. Even though a more accurate algorithm could need more computing power, it might perform better all-around in terms of accurately identifying emotions.

- 2) Feature Extraction: The process of extracting significant features from raw audio and video data, which can then be used for analysis, is a crucial stage in the emotion detection process. Performance can be substantially influenced by the feature extraction method used, as certain methods may require more computation than others.

- 3) Hardware requirements: The execution of a product may also be dependent on the hardware needed for emotion recognition. Although it can speed up calculations and boost overall performance, high-performance technology might also be more expensive and not be available in all circumstances.

- 4) Data quantity: Performance may also be driven by the magnitude of the data used for emotion recognition. Larger datasets could need more processing power and take longer to process, which could affect the system's overall implementations.

- 5) Real-time rendition: Real-time rendition is essential in some applications, such as live video or audio analysis. To ensure that emotions can be recognized and evaluated in real-time, the system should be developed to reduce processing time and latency.

#### **4.2.2.1.4 Accuracy**

The accuracy generated by our work is outperformed by any other existing models. We can recognize emotions and eye drowsiness accurately through our proposed system. Accuracy, or a system's capacity to correctly identify emotions in audio and video data, can have major consequences on real-time emotion detection. These include:

- 1) Dataset quality: The relevancy of the emotion recognition algorithm can be significantly impacted by the quality and diversity of the training dataset. The algorithm may be trained on a wide range of emotions and be better able to manage variances in audio and video data with a larger and more varied dataset.
- 2) Selection of characteristics: The accuracy of emotion detection may be impacted by the features chosen. To accurately capture the most significant aspects of audio and visual data connected to emotions, the features should be carefully selected.
- 3) Selection of the algorithm: Accuracy may be additionally determined by the algorithm adopted for emotion recognition. The choice of the best algorithm can considerably increase accuracy because different algorithms may behave differently on various datasets.
- 4) Preprocessing: The exactness of audio and video data can also be affected by preprocessing. The quality of the data can be enhanced using methods like noise reduction, filtering, and normalization, which will make it simpler for the algorithm to correctly identify emotions.
- 5) Validation: Lastly, the reliability of the emotion detection system can be helped by validating it. To make sure that the system can reliably recognize emotions in previously unobserved data, it needs to be verified on a different test dataset.

### **4.2.2.2 ORGANIZATIONAL REQUIREMENTS**

#### **4.2.2.2.1 IMPLEMENTATION REQUIREMENTS**

- Hardware Requirements:

Complex deep learning models for audio and video analysis of human emotions require powerful technology, such as GPUs, CPUs, and specialized deep learning

gear.

- **Information Gathering and Annotation:**

Deep learning models need to be trained on a lot of data. For the models to be trained properly, this data must be accurately collected and labeled.

- **Data Preprocessing:**

Before using the audio and video data to train the deep learning models, preprocessing is required. This entails transformation, feature extraction, feature normalization, and data cleaning.

- **Model choice and instruction:**

The selection of suitable deep learning models is necessary for the analysis of human emotions in audio and video. The preprocessed data must be used to train the models.

- **Performance Evaluation:**

It is necessary to assess the effectiveness of the trained models using a variety of metrics, including accuracy, precision, recall, F1-score, and others.

- **Deployment:**

The models must be installed on the production environment following successful training and evaluation. Setting up the necessary infrastructure and integrating the models with the current systems are required for this.

- **Supervision:**

In order to make sure the deployed models are working well and to spot any potential problems, they must be monitored constantly. This involves keeping an eye on the model's precision, responsiveness, and resource usage.

#### **4.2.2.2.2 ENGINEERING STANDARD REQUIREMENTS**

- **Data Caliber:**

The training and testing data must be of good quality and indicative of real-world scenarios. To enable efficient training and validation, the data must be appropriately annotated with labels and metadata.

- **Performance of Model:**

Model performance must be precisely measured and optimized in order to achieve high accuracy, low latency, and optimal resource consumption. The models must be built to withstand noise, changes in lighting, and other environmental influences.

- **Interpretation of Model:**

The models must be constructed to be understandable and interpretable. This means that the models must provide explicit explanations for their decisions and forecasts in a way that humans can understand.

- **Robustness of Model:**

Models must be resistant to adversarial attacks and other types of interference. This necessitates the implementation of proper security measures as well as robustness testing.

- **Scalability:**

Scalable models that can handle massive amounts of data and user traffic are required. This necessitates the use of efficient data processing and storage systems, as well as load balancing and other performance enhancement strategies.

- **Compliance:**

Deep learning models for audio and video analysis of human emotions must be implemented in accordance with all relevant legal and regulatory standards, such as data privacy and security rules.

- **Documentation:**

It is essential to thoroughly record the implementation in order to make future changes and maintenance simple. This involves outlining the performance metrics, training and validation processes, model architecture, data sources, and model architecture.

### 4.2.2.3 OPERATIONAL REQUIREMENTS

- Economy

There are several commercially viable use cases for this project, encompassing sectors like market research, customer service, healthcare, and entertainment. In market research, our project can assist businesses in better comprehending the emotional responses of their clients to their goods and services, allowing them to decide on product development and marketing tactics with knowledge.

- Environmental

Our project can be used in a variety of environmentally friendly applications, including energy efficiency, smart home technologies, environmental monitoring and transportation. We can contribute to the creation of a future that is more ecologically friendly and sustainable by utilizing deep learning models in these fields.

- Social

The fact that our research can be used in a variety of socially relevant use cases across a variety of sectors, including mental health, education, social services, and robotics, is one of its most significant uses. We can raise people's quality of life and forge a society that is more compassionate and receptive by utilizing deep learning models in these fields.

- Political

Our thesis has a wide range of politically relevant use cases that can be used to different fields, including law enforcement, political debate, and policymaking. Deep learning models can be used in these areas to enhance the political process, boost accountability and transparency and establish more responsive and efficient governance.

- Ethical

With the aid of our project, a number of ethical issues can be resolved. However, deep learning algorithms for audio and video analysis of human emotions can be utilized morally to support beneficial societal outcomes by upholding fairness, privacy, transparency and social responsibility.

- Health and Safety

There are a variety of applications for our project in terms of health and safety. We can enhance the outcomes of mental healthcare, reduce accidents, enhance healthcare diagnosis and treatment, and respond to emergency situations more skillfully by utilizing deep learning models in these fields.

- Sustainability

There are several applications with respect to sustainability for the use of deep learning models for audio and video analysis of human emotions. We can decrease energy use, boost transportation effectiveness, encourage trash reduction, and improve sustainable agricultural practices by utilizing deep learning models in various fields. In the transportation sector, the models we employed in our project can enhance traffic management systems to save energy consumption and boost transportation efficiency by identifying emotional cues that indicate traffic congestion.

- Legality

The audio and visual analysis of human emotions has a variety of legitimate uses. We can increase public safety, advance justice, assure regulatory compliance, and defend against criminal activity by utilizing deep learning models in various fields. However, it is crucial to make sure that these technologies are employed in accordance with moral and legal standards, and that the necessary precautions are taken to protect civil liberties and privacy. In cyber security, the models invoked in our research can assist in enhancing cyber security and defending against unlawful activities by identifying emotional cues that may point to a potential cyber-attack or security breach.

- Inspectability

In terms of inspectability, our project possesses several applications. We can encourage accountability, openness, and ethical usage of these technologies by making sure that the model's decision-making process can be looked at and understood. In education sector, the deep learning models used by us can help educators and teachers deliver individualized instruction. To make sure that the model is not supporting prejudices or inequalities in the educational system, it is crucial that the model's decision-making process can be examined.



## **4.2.3 SYSTEM REQUIREMENTS**

### **4.2.3.1 HARDWARE REQUIREMENTS**

- GPU
- CPU
- RAM
- Storage and cooling.
- These components offer a solid starting point, but the precise hardware requirements will vary depending on the implementation details and the size of the dataset being examined

### **4.2.3.2 SOFTWARE REQUIREMENTS**

- Tensorflow and Keras Frameworks
- Python programming language
- Windows Operating System v1
- Kaggle Notebook IDE
- Audio Analysis Library- Librosa
- Video Analysis Library-Facial Expression Recognizer(FER)

## 5. RESULTS AND DISCUSSIONS

### Video

#### Average Of emotions from the Input Video

		1 to / of / entries	Filter	
index	Human Emotions	Emotion Value from the Video		
0	Angry	3.739999999999983		
1	Disgust	0.0		
2	Fear	40.50000000000009		
3	Happy	303.9599999999998		
4	Sad	9.919999999999945		
5	Surprise	8.12999999999994		
6	Neutral	114.0799999999998		

Fig 5.1: It is showing the emotion value of the face during that period of time that is recorded

#### Graph of emotions

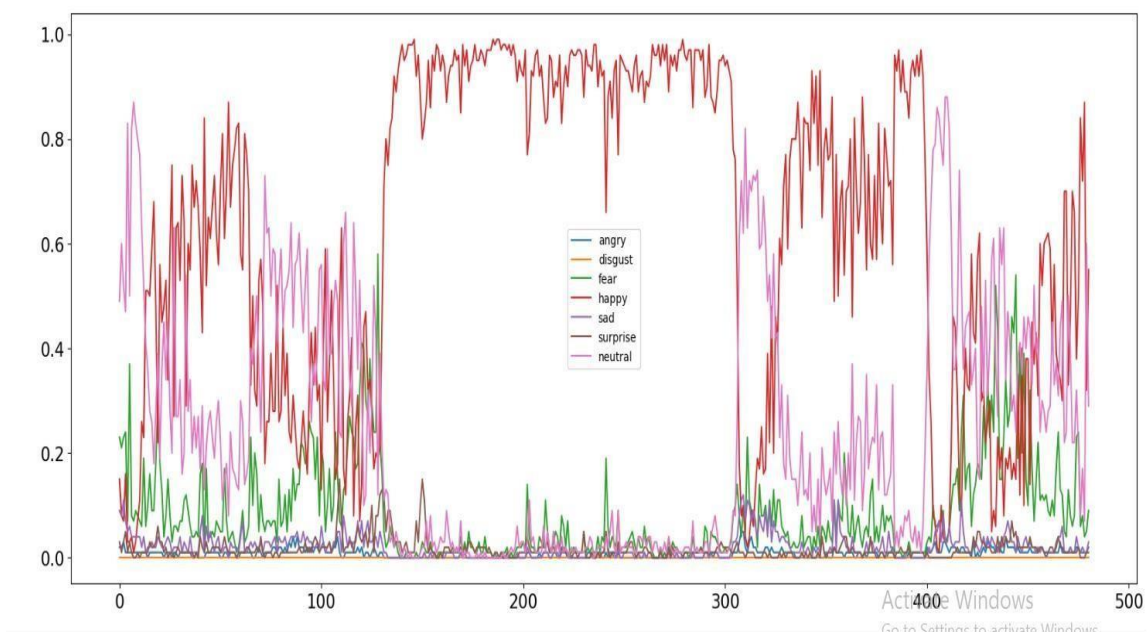
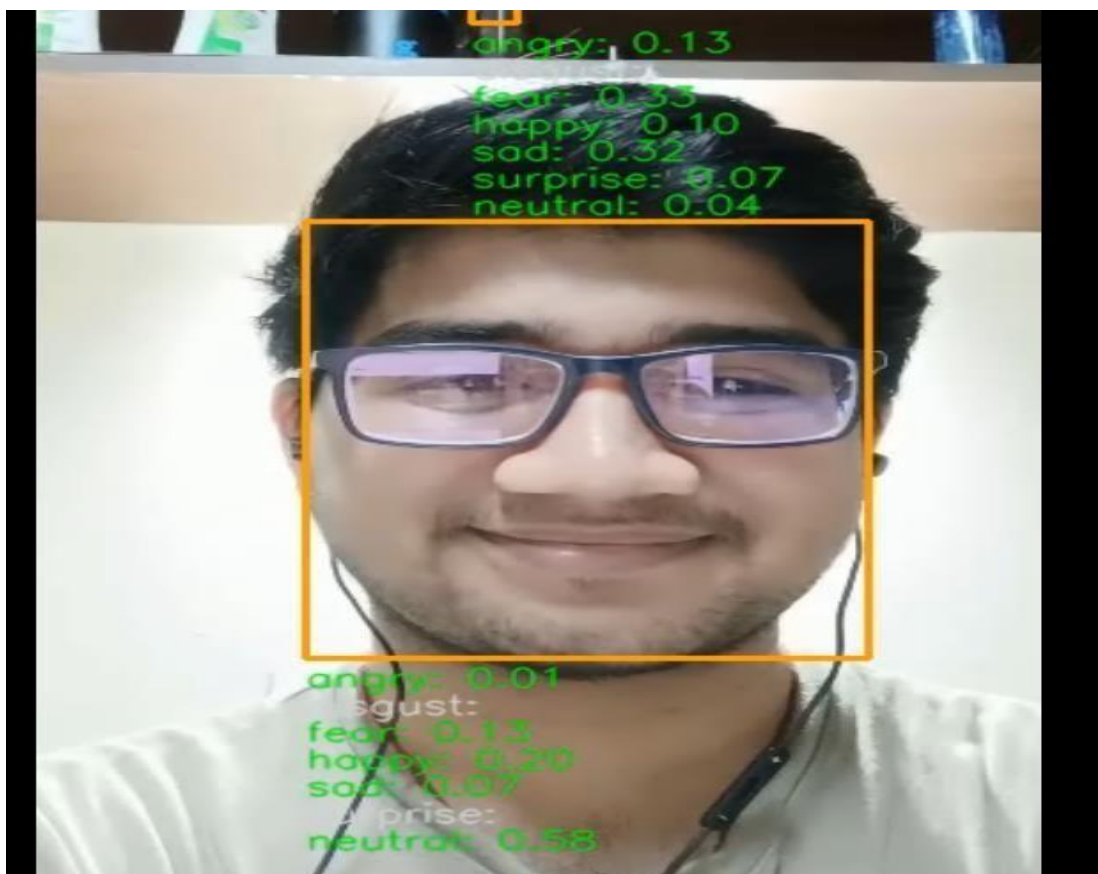
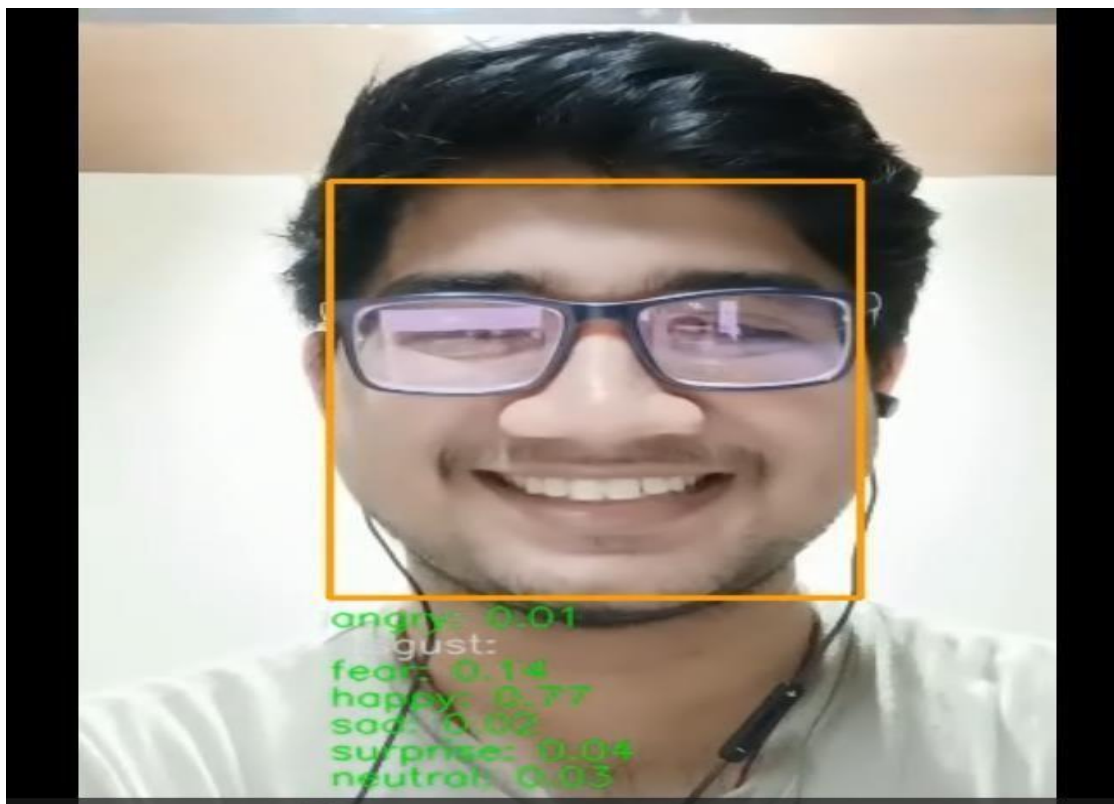


Fig 5.2: This is the graph of the emotion analyzed through the input video

## Output Result



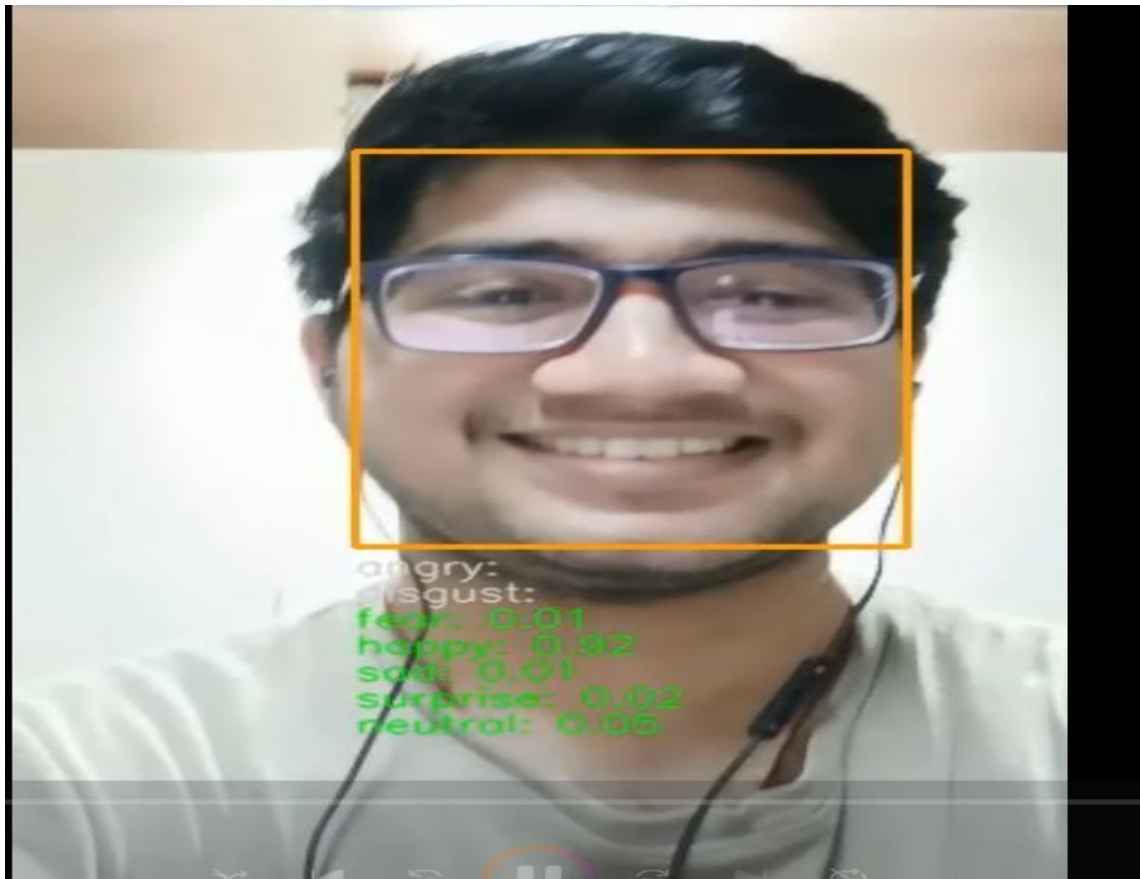
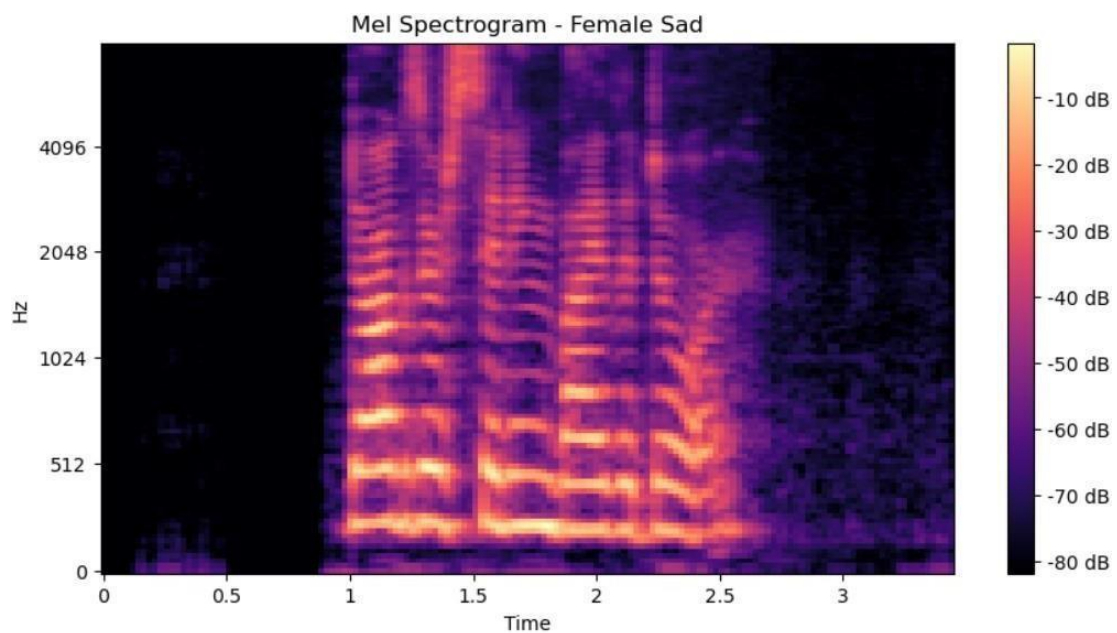


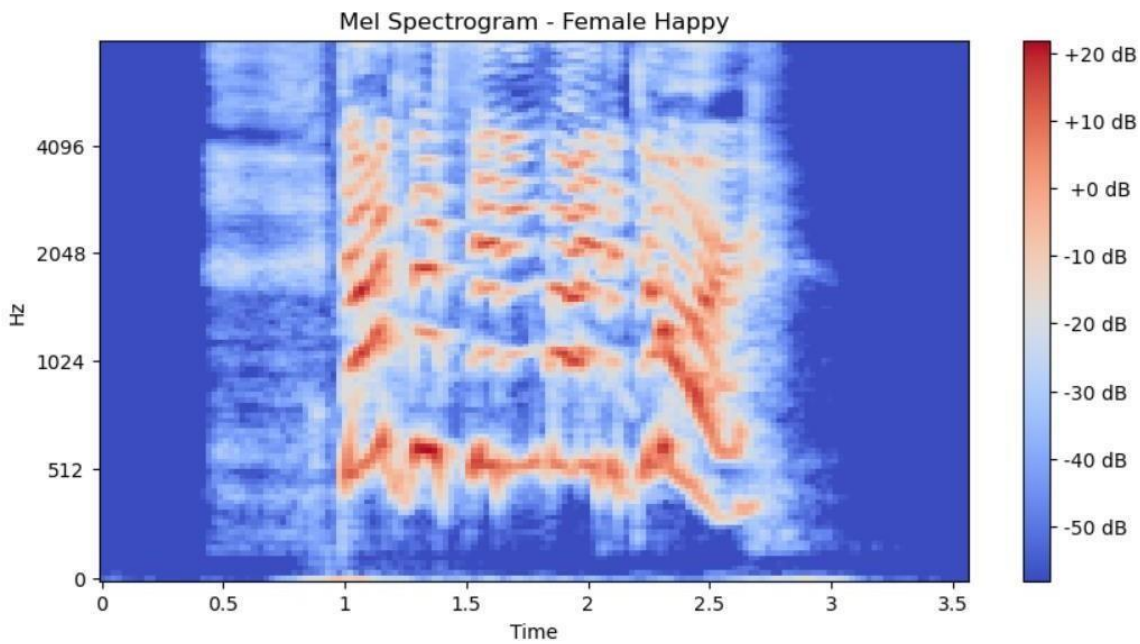
Fig 5.3: This is the emotional value that has appeared on the screen during each expression I am giving

## Audio

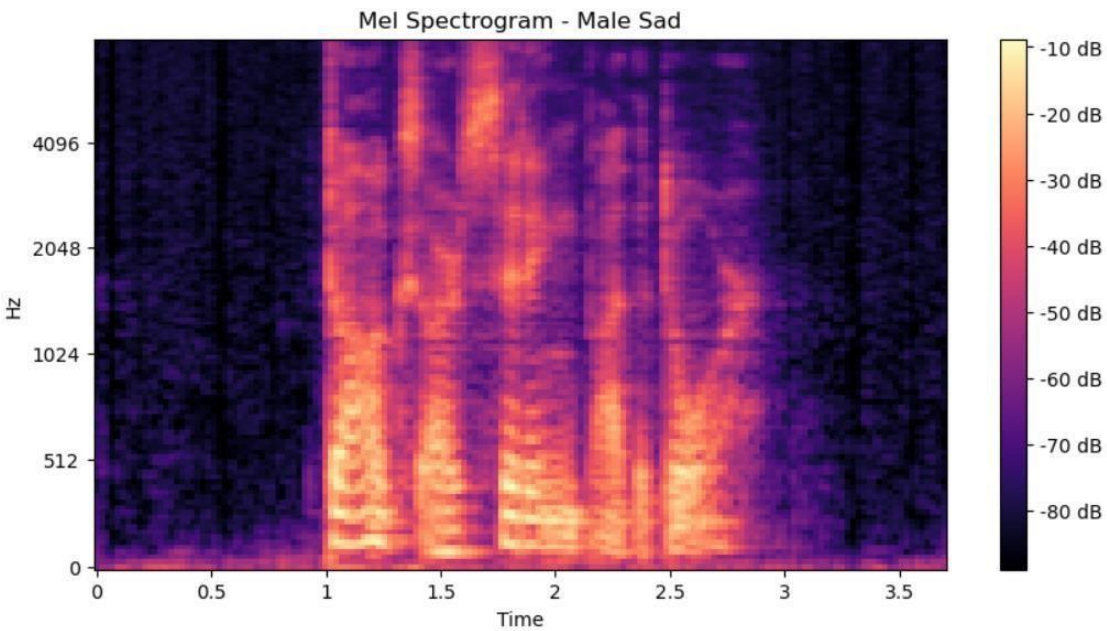
### Female Sad Mel Spectrogram



Female Happy Mel Spectrogram

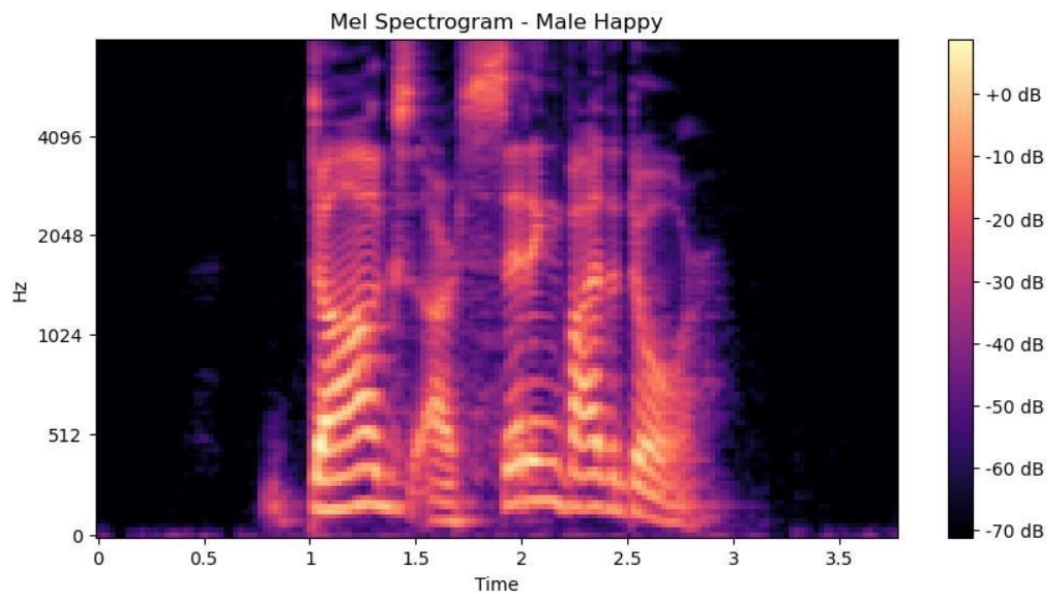


Male Sad Mel Spectrogram





## Male Happy Mel Spectrogram



The above figures are Mel Spectrogram for specific emotions we have used

## Efficiency Using MLP Model

Training set score: 0.922  
Test set score: 0.590

## Efficiency Using LSTM Model

23/23 [=====] - 0s 11ms/step - loss: 0.0462 - accuracy: 0.5833  
Accuracy of our model on test data : 58.3333134651184 %

## Efficiency Using KNN Model

Training set score: 0.533  
Test set score: 0.324

## Efficiency Using GRU Model

23/23 [=====] - 0s 12ms/step - loss: 0.0642 - accuracy: 0.2000  
Accuracy of our model on test data : 20.00000298023224 %

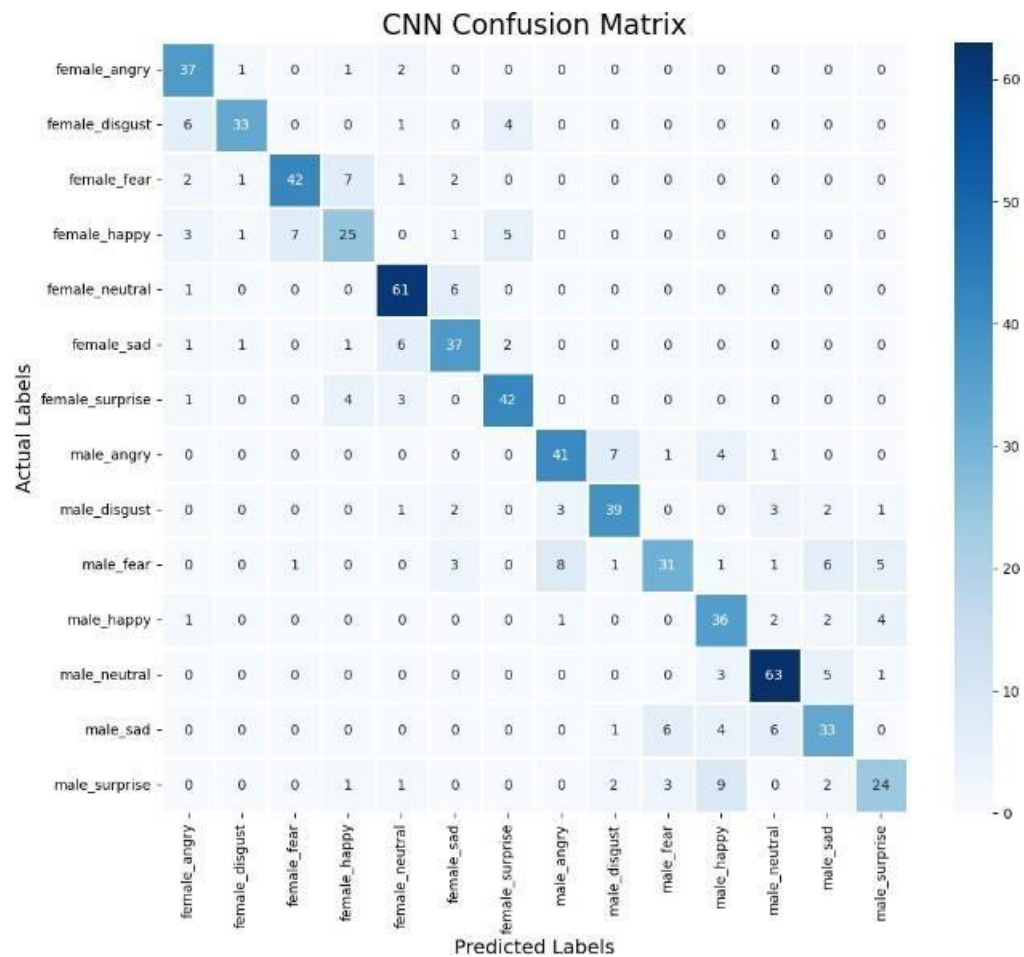
## Efficiency Using Decision Tree Model

Training set score: 1.000  
Test set score: 0.410

Efficiency Using CNN Model

23/23 [=====] - 2s 92ms/step - loss: 1.7151 - accuracy: 0.7556  
Accuracy of our model on test data : 75.5555701255798 %

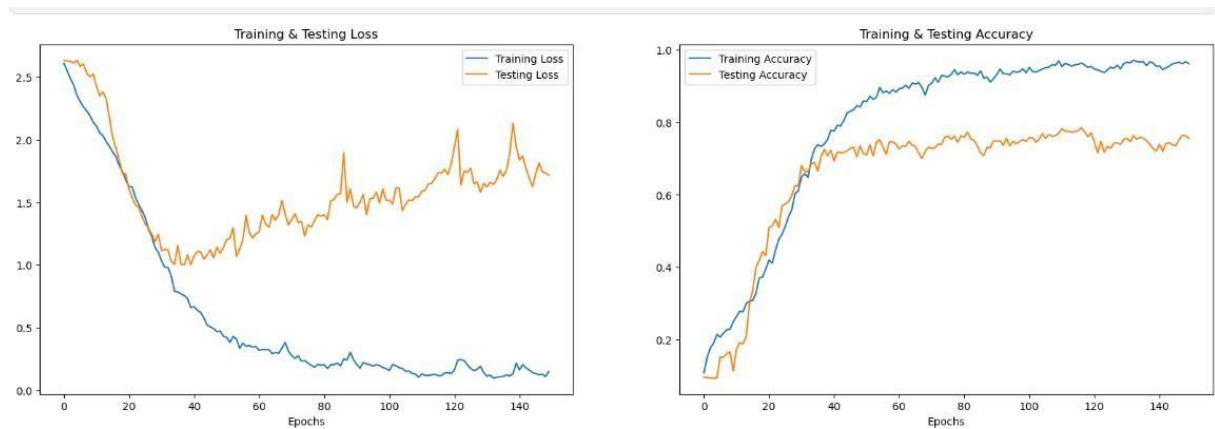
CNN Confusion Matrix



Since CNN model has the highest efficiency so this model has been used for our live demo

By analyzing the confusion matrix, you can gain insights into the model's performance for each class. It allows you to assess the accuracy, precision, recall, and F1 score for individual classes and overall performance.

## Loss Accuracy Plot using CNN



The plot below displays the loss and accuracy values for the CNN model during the training and testing phases

## Input Audio Frequency Graph

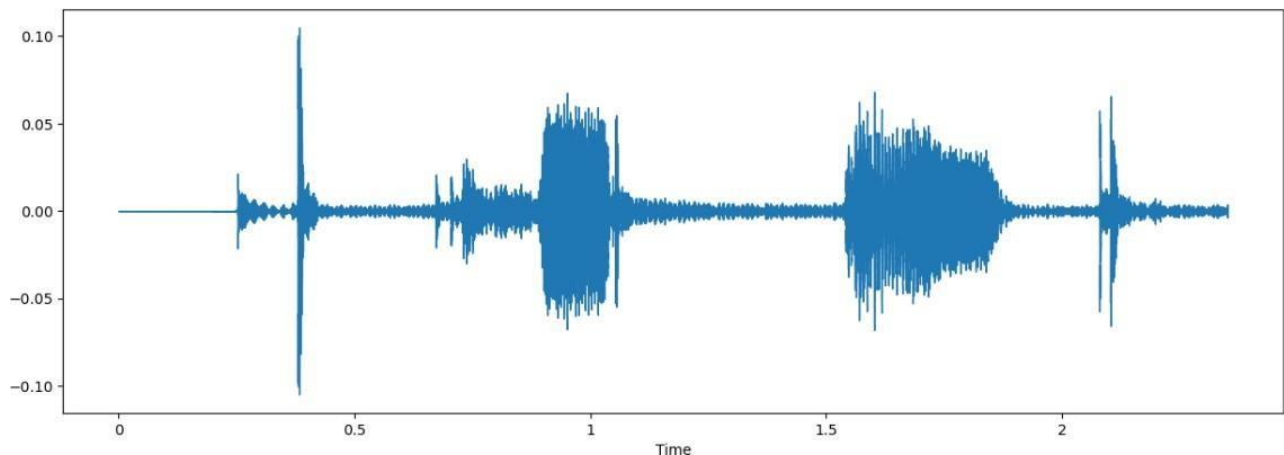


Fig 5.4: The following is a frequency graph representing the input audio recording that has been added.

## Final Output

```
array([[ 'female_sad' ]], dtype=object)
```

The result of the audio recording that was previously submitted has been provided.



## 6. REFERENCES

- [1] Ouellet, Sébastien. "Real time emotion recognition for gaming using deep convolutional network features." arXiv preprint arXiv:1408.3750(2014).
- [2] Pranav, E., Suraj Kamal, C. Satheesh Chandran, and M. H. Supriya. "Facial emotion recognition using deep convolutional neural network." In 2020 6<sup>th</sup> International conference on advanced computing and communication Systems(ICACCS),pp.317-320.IEEE,2020.
- [3] Abdullah,SharmeenM.SaleemAbdullah,SiddeeqY.AmeenAmeen,MohammedAMSadeq, and Subhi Zeebaree. "Multimodal emotion recognition using deep learning."Journal of Applied Science and Technology Trends 2, no.02 (2021): 52-58.
- [4] Khanzada,Amil,CharlesBai,andFerhatTurkerCelepcikay."Facial expression recognition with deep learning."arXiv preprint arXiv:2004.11823 (2020).
- [5] Khalil, Ruhul Amin, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan,MohammadHaseebZafar,andThamerAlhussain."Speech emotion recognition using deep learning techniques: Areview."IEEEAccess 7 (2019): 117327-117345.
- [6] ElMettiti,Abderrahmane,MohammedOumsis,AbdellahChehri,andRachidSaadane. "Real-Time Emotion Recognition Using Deep Learning Algorithms." In 2022 IEEE 96<sup>th</sup> Vehicular Technology Conference(VTC2022-Fall), pp.1-5.IEEE,2022.
- [7] Li, Junnan, and Edmund Y. Lam. "Facial expression recognition using deep neural networks."In2015 IEEE International Conference on Imaging Systems and Techniques(IST),pp. 1-6.IEEE, 2015.
- [8] Sinha, Avigyan, and R.P.Aneesh. "Real time facial emotion recognition using deep learning." International Journal of Innovations and Implementations in Engineering 1(2019).

[9] Jermittiparsert, Kittisak, Abdurrahman Abdurrahman, ParinyaSiriattakul, Ludmila A.Sundeeva, WahidahHashim, RobbiRahim, and Andino Maseleno. "Pattern recognition and features selection for speech emotion recognition model using deep learning." *International Journal of Speech Technology* 23 (2020): 799-806.

[10] Wu, Ching-Da, and Li HengChen. "Facial emotion recognition using deep learning." *Ar Xivpreprint arXiv:1910.11113* (2019).

[11] Jaymon, Noel, Sushma Nagdeote, Aayush Yadav, and Ryan Rodrigues. "Real time emotion detection using deep learning." In *2021 International conference on advances in electrical, computing, communication and sustainable technologies (ICAECT)*, pp. 1-7. IEEE, 2021.

- [12] Sang, Dinh Viet, and Nguyen Van Dat. "Facial expression recognition using deep convolutional neural networks. " In 2017 9<sup>th</sup> International Conference on Knowledge and Systems Engineering (KSE), pp. 130-135.IEEE,2017.
- [13] Song,Inchul,Hyun- JunKim ,and Paul Barom Jeon."Deep learning for real-time robust facial expression recognition on a smartphone." In 2014 IEEE International Conference on Consumer Electronics (ICCE),pp.564-567.IEEE,2014.
- [14] Fathallah, Abir, LotfiAbdi, and AliDouik. "Facial expression recognition via deep learning." In 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications(AICCSA), pp. 745-750.IEEE,2017.
- [15] Zhang, Qiang, Xianxiang Chen, Qingyuan Zhan, Ting Yang, and Shanhong Xia."Respiration based emotion recognition with deep learning." Computers in Industry92(2017): 84-90.