Ain Shams University
Faculty of Computer and Information Sciences
BIOINFORMATICS PROGRAM

Psychic Detective

By
Sarah Ahmed Mohamed Mahmoud
Hagar Hossam El-Dein Saed Ewees
Mostafa Hasanen Ahmed Hasanen
Nagham Reda Ibrahim Hassan
Nouran Sayed Hamza Hassan
Shimaa Hamdy Labib Yousef

Under Supervision of
Dr. Mahmoud Mounir
Assistant Professor in Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University

TA. Esraa Hamdi
Assistant Lecturer in Bioinformatics Department
Faculty of Computer and Information Sciences,
Ain Shams University

Acknowledgement

First of all, we would like to thank Allah for being with us in all our ups and downs and for giving us the strength to move forward along this journey of hard work and in life generally, we are grateful to our parents and family who were always providing help and support throughout the whole years of study, we hope we can give that back to them, Also We want to express our gratitude for our supervisor DR. Mahmoud Mounir, he has always been supportive to us through the whole work, we would like to also thank our dear teaching assistant TA. Esraa Hamdi, finally we thank all our fellow students in all departments for this heartwarming, supportive and competitive community that was like a huge family to us.

Abstract

Who among us has gone through a difficult period in his life and thought that he needed a psychiatrist?
In our project, we aim to break the barrier of fear from the idea of society towards mental disorder through questions and answers through which information about the user can be collected and enable us to make a preliminary diagnosis of the person and find out if he needs a psychiatrist or not.
We design and train a deep neural network to perform this task using ~60,000 of natural images and records of people speaking from Internet. During training, our model learns audio with providing an accuracy of 73.4%, image-face correlations with accuracy of 95.5% to detect emotions under Human Basic Emotions.
Seeing in our project how emotions perform while preserving aging accuracy so that we ensure that the generated emotion desired aging effects while simultaneously keeping personalized properties stable.

Table of Contents

List of Figures

List of Abbreviations

NAMI	National Alliance on Mental Illness
WHO	World Health Organization
VED	Video Emotion Detection
Crema-D	Crows-sourced Emotional Multimodal Actors Dataset
Savee	Surrey Audio-Visual Expressed Emotion
Ravdess	Ryerson Audio Visual Database of Emotional Speech and Song
Tess	Toronto Emotional Speech Set
RMS	Root Mean Square
MFCC	Mel-frequency cepstral coefficients
ST	Sensibility Technology
VEA	Voice Emotion Analysis
PST	Psychoanalysis System Technology
KNN	K-Nearest Neighbor
BPD	Borderline personality disorder
LSTM	Long Short-term Memory
SVM	Support Vector Machine
CNN	Conventional Neural Network
GAN	Generative Adversarial Network

List of Tables

Chapter 1

Introduction

1.1 Motivation

- Mental health is just as important as physical health but in most of Middle East it is taken as an unnecessary expensive sessions and waste of money and time, resulting in reaching higher levels of depression, suicidal cases, ignoring mental issues and suppressing any emotions.

- According to statistics conducted by NAMI that one from each four people suffer from a mental disorder, and the high number of suicides nowadays and lack of awareness of the mental health problems affecting majority of population and amplifying of the stigma associated with mental illness.

- Spread awareness of early warning signs:
    1. Mood swings or constantly feeling low
    2. Thinking life is not worth living
    3. Disturbed sleep, either not getting enough or sleeping too much
    4. Eating less than normal or overeating, perhaps losing or gaining weight

Figure 1.1: early warning signs

1.2 Problem Definition

1. Lack of psychotherapy seeks when in need; considering it as a shame or catalyst for inducing the sadness (Not knowing feeling emotions is cure not disease), specifically in Middle East: People consider it as fraud or not worth taking the sessions; thinking they can treat it internally or by ignoring it or misunderstand and relate it to different beliefs.

2.Unawareness of your emotions leads to inability to access feelings, struggle to understand others, can't control emotions, feel numb inside, or feel so emotional and need to try to escape feelings by wrong sort of ways.

3. Mental health problems affect around one in four people in any given year.

4. Medical conditions are accurately and effectively treated best when they are diagnosed very early, when your mental illness is in its early stage, you can easily prevent the symptoms from worsening.

5. Recently, the quarantine has affected the health of many people, whether mental or physical, so many people are wondering whether they are in a mental health condition or have they already been affected by the period of home quarantine.

1.3 Objective

- Construct system aims to help users keep track of their mental health status.

- By building an emotion detection system consist of three levels:

    1. Questionnaire: By responding to determined questions from a psychological doctor about state of mind and mentality information, and by capturing the provided answers and the way of response in the use of words to determine emotions of the user.
    2. Voice Recognition: tone and pitch in voice mostly reflect underlying emotion that can be classified to different types and used in more specifying psychic cases.
    3. Visual Response: by capturing facial state and body language.

    - Through the previous stages, the user's feelings can be determined from human basic emotion

Figure 1.2: Basic Human Emotion

- Design a deep neural network and train it to perform this task using millions of natural voices and images, it's used in several fields, e.g., detect an approximate emotion.

- Get high-accuracy results and close to reality.

- Developing Flexible UI for easy usage.

1.4Time Plane

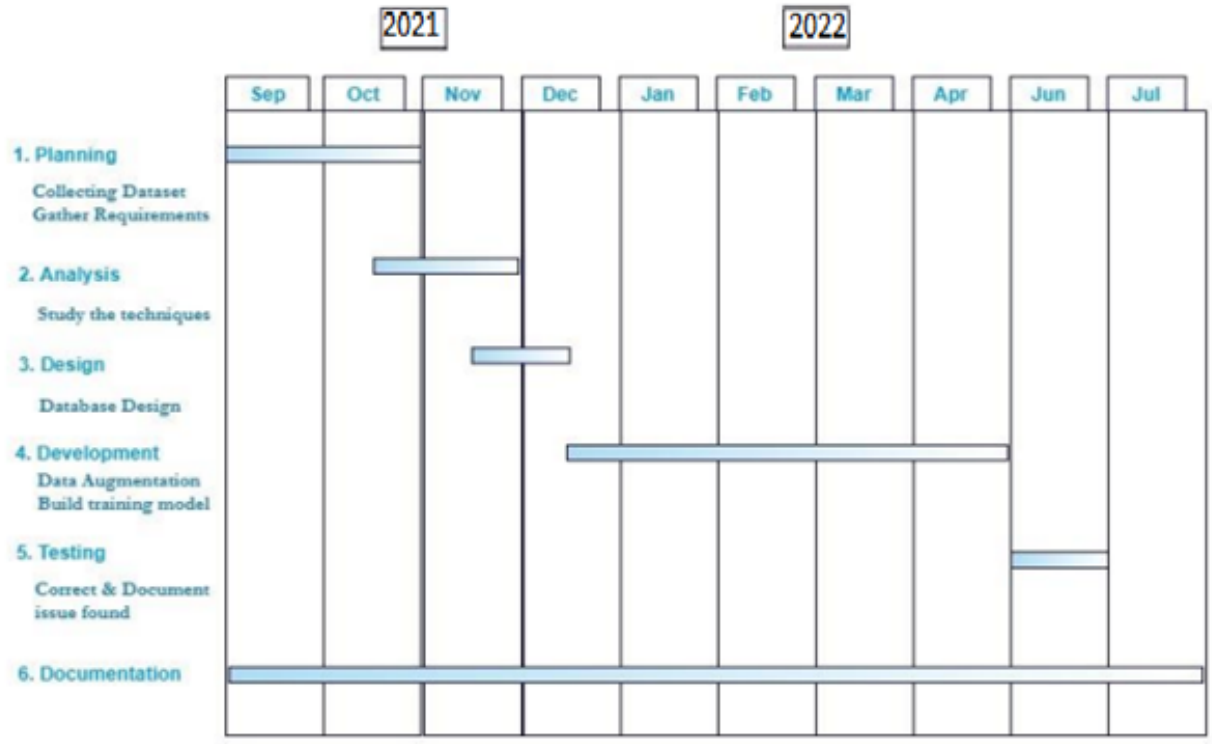


Figure 1.3 Time Plan

1.5 Document Organization

In **chapter 2**, the background of the project was discussed, recent related works were mentioned, and datasets used for each one. Comparison between results of different approaches done by others based on used algorithms, datasets, and overall accuracy.

In **chapter 3**, the system architecture was discussed, also system users were mentioned. The use case diagram is included as well as its fully dressed use cases. The class diagram and explanation of each class and sequence diagrams for some of the important use cases in the use case diagram are included.

In **chapter 4**, the algorithms used in each phase were discussed as well as the pseudo-code for each major function in the system. UI implementation and testing stages were discussed.

In **chapter 5**, includes the user manual, essential information for the user to make full use of the system, screenshots are also attached from the app and an installation guide.

In **chapter 6**, conclusion and future work and improvements were discussed

Chapter 2

Background

- **2.1 Overview**
  - Mental health is a state of well-being in which a person understands his or her own abilities, can cope with the normal stresses of life, can work productively and fruitfully, and is able to contribute to his or her community (WHO) Mental-health is affected by past traumas, daily interactions, relationships, physical health, and natural causes.

  - The Perspectives of Psychology have become more dominant and influential approaches of biological, psychodynamic, behavioral, cognitive, and humanistic evolved from Structuralism and functionalism.

- **2.2 Scientific background**
  - Real Time Facial Expression Recognition App Development on Mobile Phones (Published in 2016)

    - This paper discusses Facial Expression Recognition System which performs facial expression analysis in a near real time from a live web cam feed. Primary objectives were to get results in a near real time with light invariant, person independent and pose invariant way. The system is composed of two different entities trainer and evaluator. Each frame of video feed is passed through a series of steps including here classifiers, skin detection, feature extraction, feature points tracking, creating a learned Support Vector Machine model to classify emotions to achieve a tradeoff between accuracy and result rate. A processing time of 100-120 m/s per 10 frames was achieved with accuracy of around 60%. We measure our accuracy in terms of variety of interaction and classification scenarios. We conclude by discussing relevance of our work to human computer interaction and exploring further measures that can be taken.

- KNN → It is a supervised machine learning algorithm. The algorithm can be used to solve both classification and regression problem statements.

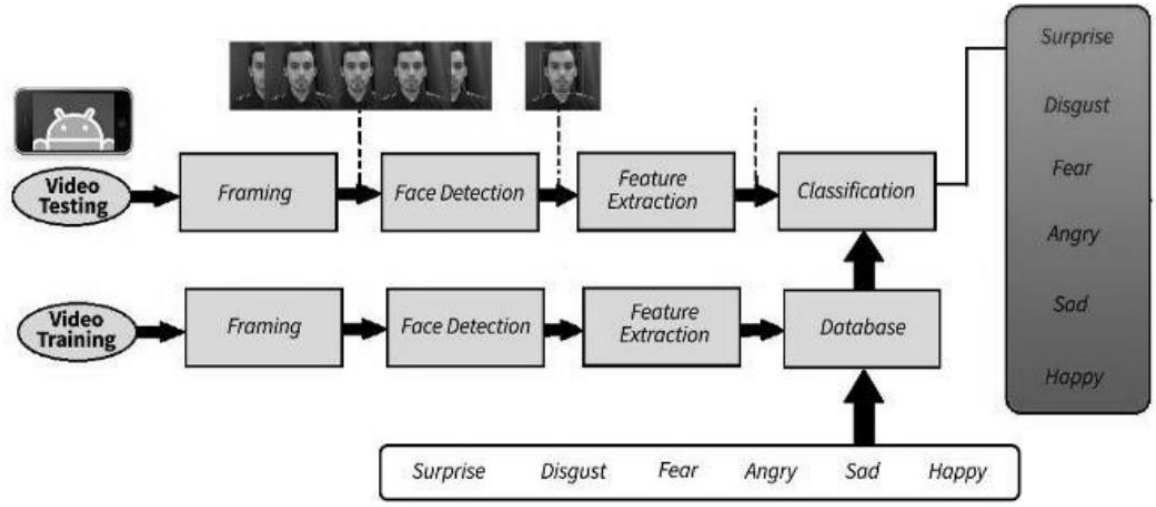


Figure 2.1: overview structure of the facial expression recognition

➢ Emotion Recognition Using Mobile Phones (published in 2016)

- This paper presented a machine learning approach for emotion recognition using a mobile phone soft keyboard. The keyboard records the user's typing behavior that includes texting speed and time between presses and shaking as measured through the built-in accelerometer. The keyboard dynamically uses the J48 machine learning algorithm to classify the user's current mood. The system also sends anonymized user data to a server that can be publicly accessed to view demographic information. The demographic data could be used by researchers in various fields and disciplines. The system demonstrates that it is possible to enable emotion recognition on mobile phones using built-in sensors. The system also does so in an application independent manner where any mobile application using a keyboard for input can use the proposed service. The emotion detection service can be used in many domains, such as healthcare, social media, trading, etc.

- J48→ decision tree that splits the data into different subsets Multi-response nearest neighbor approach, where the distance between two feature vectors is calculated and a class is assigned based on the nearest neighbor.

- linear regression→ Classification possibilities are converted into binary and a regression model is created for each possible class.
- Lazy IBK→ nearest neighbor approach, where the distance between two feature vectors is calculated and a class is assigned based on the nearest neighbor
- SVM→ Creates a hyper-plane separating the various emotions

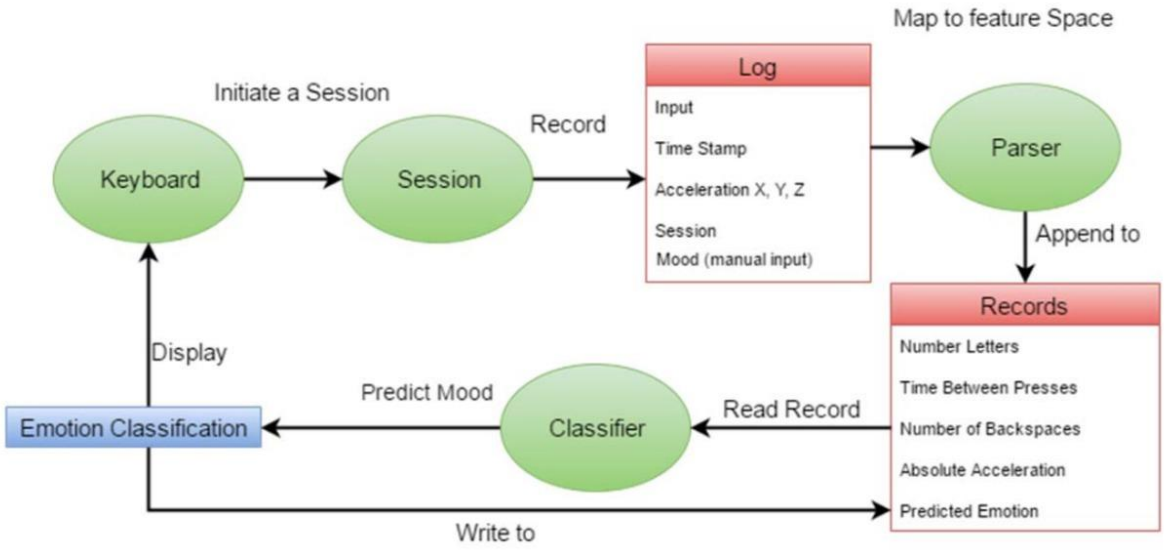


Figure2.2: data capture and prediction

- **2.3 Description of existing similar systems.**

  - MIMOSYS for voice-based health monitoring

    - In 2016 developed APP MIMOSYS measures the prosodic parameters of the voice and detects emotional components contained in the voice. Then, the degree of stress or depressive symptoms is estimated from changes in calculated emotional components using ST and VEA

      - ST→ is the technology which allows you to analyze human's complex feelings automatically from voice regardless of a word itself or individual differences.
        By leveraging ST, human's complex feelings can be quantitatively visualized by colors.

      - VEA→ refers to the use of various methods to analyze vocal behavior as a marker of affect (e.g., emotions, moods, and stress), focusing on the nonverbal aspects of speech.

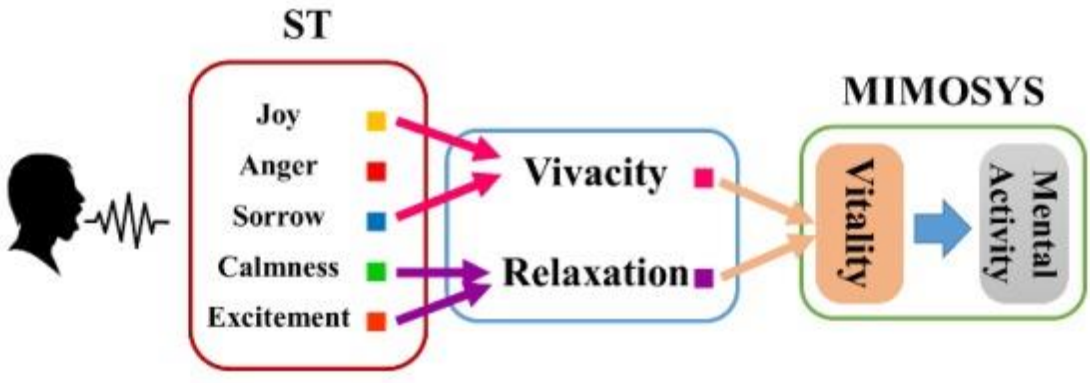


Figure 2.3: MIMOSYS

- MindTime (Published in 2020)

  - This paper introduces the MindTime system, which aims to detect Borderline personality disorder symptoms and signs of Self-harm ideation. MindTime is a personal diary mobile app. It allows users to make notes of daily events, experiences, thoughts, and feelings. The system features include password protection, mood tracking, text journals, voice recording (with speech to text), and video recording. Once the user adds a diary input, it is analyzed to detect if there are signs of BPD symptoms. We investigated different classifiers to extracts features from the stored diaries based on the input type text and video. The used classifiers were Naive Bayes, SVM, KNN, and finally LSTM. For text, the result showed that SVM and LSTM classifiers achieved the best user accuracy of 90.1% and 91%. As for audiovisual input CNN achieved user accuracy of 65%.

    - LSTM → text classification
    - CNN → is a class of artificial neural network, most applied to analyze visual imagery (for video)

      Borderline personality disorder (BPD) is a mental illness characterized by a long-term pattern of unstable relationships and strong emotional reactions.

      Those affected often engage in self-harm and other dangerous behavior.

- ## 2.4 Related Work

Paper/Application	Author	Year	Methods	Accuracy
MIMOSYS	PST INC	2016/2021	Psychoanalysis System Technology , Sensibility Technology/ Voice Emotion Analysis	+80%
Detecting Depression Voice Signal by Chatbots	Alexandros Roniotis& Manolis Tsiknaki	2020	Beck Depression Index & Hamilton Rating Scale for Depression	59 %→ depressed 87% → Non-depressed
MindTime	MIU	2020	Support Vector Machine LSTM Convolutional Neural Network	90%→text 65%→video
Paper/Application	**Author**	**Year**	**Methods**	**Accuracy**
WYSA	Becky inkster	2018	Experience Sampling Method	62.1%
Real Time Facial Expression Recognition App Development on Mobile Phones	Brunel University London	2016	KNN	89.5%
Emotion Recognition Using Mobile Phones	American University of Sharjah	2016	Naïve Bayes, J48 Multi-response linear regression Lazy IBK and SVM	90%

Table 1: Related works.

Chapter 3

Analysis and Design

3.1 System Overview

3.1.1 System Architecture

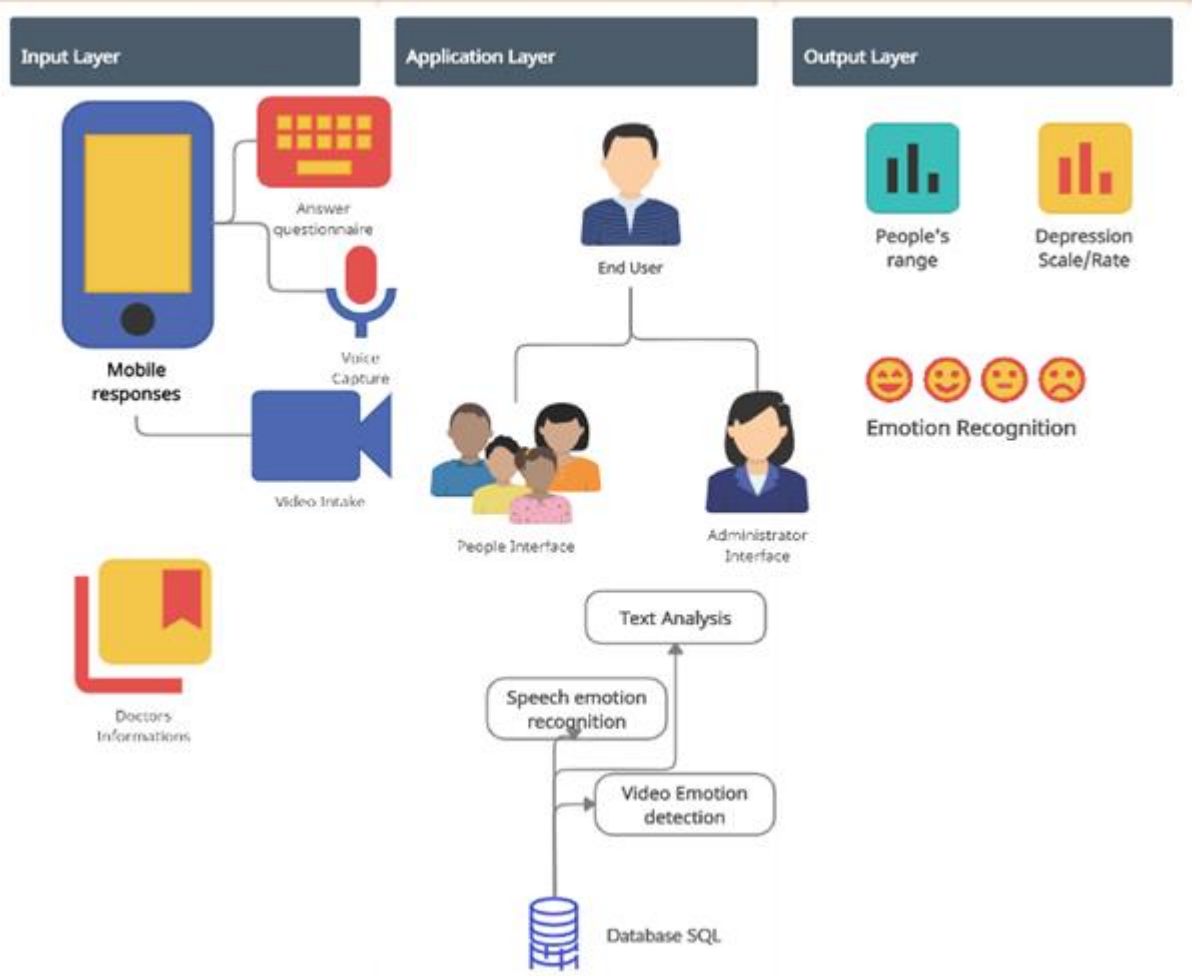


Figure 3.1: System Architecture

- The system consists of three layers
  - **Input layer:** include the three types of inputs that the user can entered (text, voice, and video).

  - **Application layer:** include types of users can entered the system (Normal user and Admin), and include models used in the system (speech emotion recognition, video emotion detection and text analysis).

  - **Output layer:** include the results that required (Depression rate and emotion recognition).

3.1.2 Functional Requirements

1. Login

- Description / Action
  - The first interface that appears when customer opens the application that check if the user already has an account or not, by checking mail and password if they exist in database.

- Requirements/ Inputs
  - Mail and password.
  - Entered Data by current user.

- Pre-condition
  - User must enter name, email, password.

- Post-condition
  - User has all access of functions depending on subscription status.
  - Go to home page.

- Output
  - Message with “Login Successfully”.

2. **Allow profile access**

   - Description/ Action
     - Allow the user to access the profile and modify its data if needed.

   - Requirements/ Inputs
     - Mail and password.
     - Entered Data by current user.

   - Pre-condition
     - User must enter correct information (Email , password)site.

   - Post-condition
     - The user answers specific questions or enters a picture or voice to express his feelings that are determined later.

   - Output
     - Have no output.

3. **Show emotion**

   - Description/ Action
     - The user answers specific questions or enters a picture or voice to express his feelings that are determined later.

   - Requirements/ Inputs
     - no input.

   - Pre-condition
     - User must enter Person's voice or image.

   - Post-condition
     - Predict user emotional.

   - Output
     - User Emotions

4. **Predict emotion**

   - Description/ Action

- It’s Deep Learning of Neural Network Algorithm using sequential model.

➢ Requirements/ Inputs
- no input.

➢ Pre-condition
- User must enter show emotion function.

➢ Post-condition
- You should still on the site until final emotion appears.

➢ Output
- specific feeling

5. Recommend doctors

➢ Description/ Action
- After predict emotion of user app can recommend doctor

➢ Requirements/ Inputs
- no input.

➢ Pre-condition
- No conditions.

➢ Post-condition
- No condition.

➢ Output
- Phone Number of doctor.

6. Load and update database

➢ Description/ Action
- admin update and load data in system.

➢ Requirements/ Inputs
- consistent data.

- Pre-condition
  - admin must enter right email and password for security .

- Post-condition
  - No conditions

- Output
  - Have no output.

7. Add doctor contact

- Description/ Action
  - Add doctor data.

- Requirements/ Inputs
  - Doctor data (clinic address _ name _ number _ email).

- Pre-condition
  - No condition.

- Post-condition
  - No condition.

- Output
  - Have no output.

8. Answer question

- Description/ Action
  - user answer group of specific questions .

- Requirements/ Inputs
  - answer of questions range from (strongly agree strongly disagree).

- Pre-condition
  - user enters correct data.

- Post-condition
  - show answers to questions.

- Output
  - Have no output.

9. show questionnaire results

- Description/ Action
  - user can show result of the questionnaire.

- Requirements/ Inputs
  - No input.

- Pre-condition
  - User must answer all questions.

- Post-condition
  - To be more accurate user enter his voice and image.

- Output
  - result and graph showing the user the extent of psychological health compared to other users of the program.

3.1.3 Nonfunctional Requirements

1. Product requirements

1.1. Security requirements

- Whatever the user's result is, it will be confidential and will not be disclosed to anyone but the owner of the result only.
- No data will be obtained from the user, whether it is a photo or video, or his answers to questions, without his prior permission.

1.2. Efficiency requirements

- Performance:
  - Time refresh does not exceed 30 seconds.
  - Response time does not exceed 10 seconds.

1.3. Usability requirements

- Friendly graphically user interface with pastel colors.

2. Organizational requirements

2.1. Development requirements

- Using java and python programming language.
- Using Android studio.
- Using PyCharm IDE.

2.2. Operational requirements

- Android application.

3.1.4 System Users

A. Intended Users:

- A person who may suffer from a mental disorder
- Normal user
- Doctor

B. User Characteristics

- No experience is required.

3.2 System Analysis & Design

3.2.1 Use Case Diagram

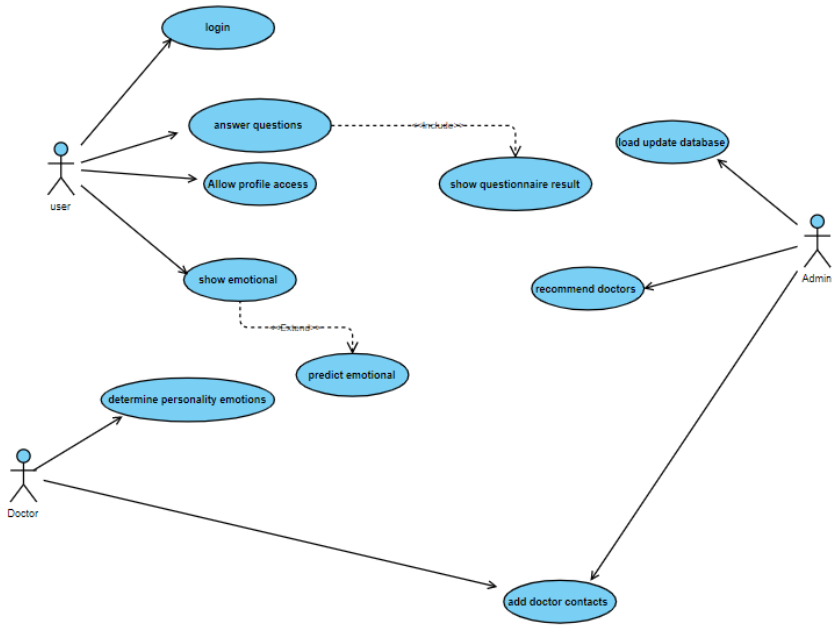


Figure 3.2: Use Case Diagram

Title	Login.
Primary Actor	The User.
Priority	Medium.
Statues	Finished.

Table 2:login

Title	Allow profile access
Primary Actor	The User.
Priority	Medium.
Statues	Finished.

Table 3:allow profile access

Title	Show emotion
Primary Actor	The User invoke it after entering a voice or image.
Priority	High.
Statues	Finished.

Table 4:show emotion

Title	predict emotion .
Primary Actor	Show emotional function
Priority	High.
Statues	Finished

Table 5: predict emotion

Title	Recommend doctor .
Primary Actor	Show emotion enter show emotion function .
Priority	Medium.
Statues	Finished.

Table 6: Recommend Doctor

Title	Answer questions
Primary Actor	The user
Priority	high.
Frequency of use	Every Time the Application is used.
Statues	Finished.

Table 7: Answer questions

Title	Show questionnaire results
Primary Actor	Answer question function
Priority	high.
Statues	Finished.

Table 8: Show questionnaire results

Title	Load and update database
Description	Process dataset.
Primary Actor	Admin.
Priority	High.
Statues	Finished.

Table 9: Load and update database

Title	Add doctor contact
Primary Actor	Admin and doctor.
Priority	Medium.
Statues	Finished.

Table 10: Add Doctor contact

3.2.2 Class Diagram

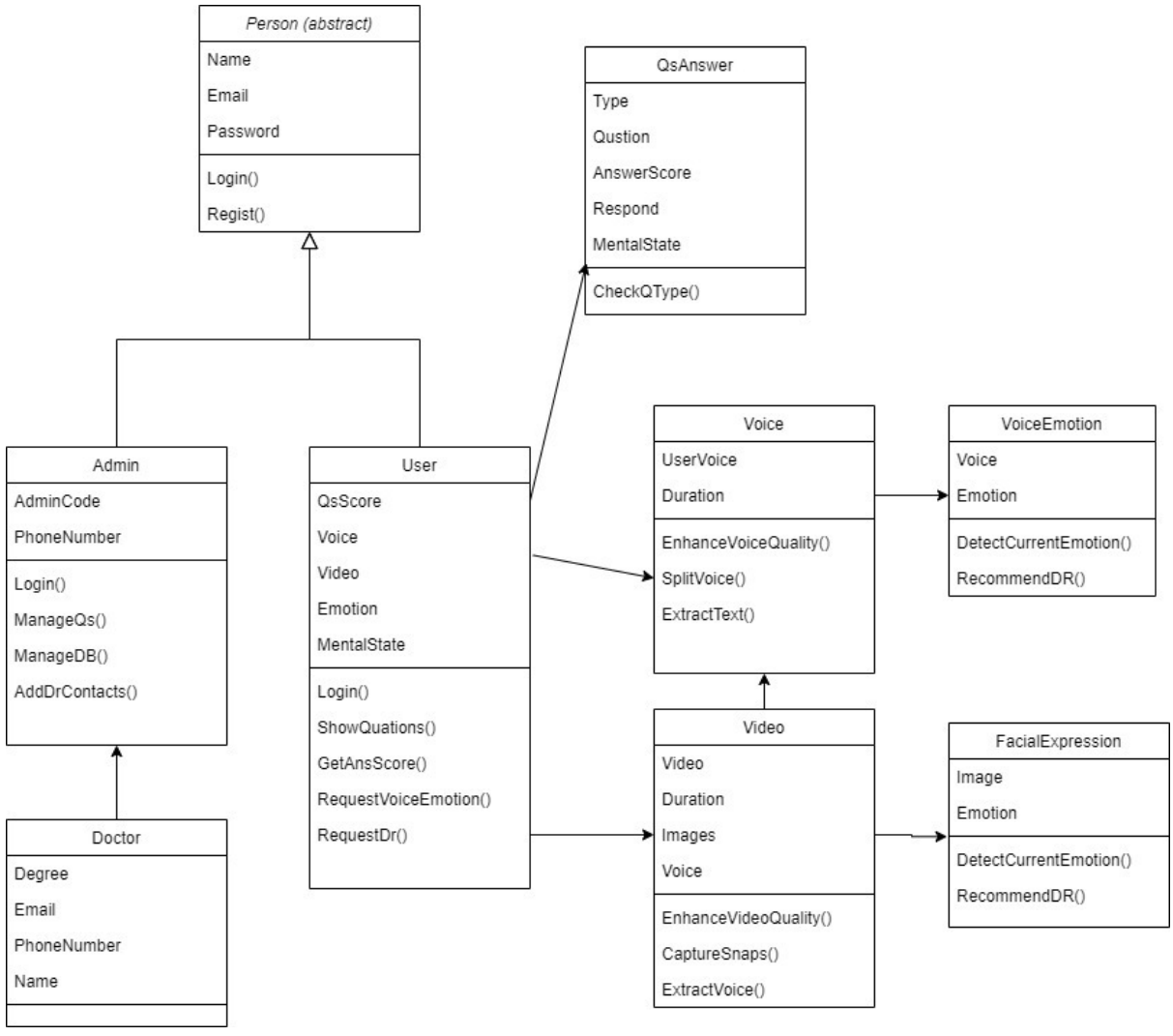


Figure3.3Class Diagram

The description of the class diagram:

Person	Abstract basic superclass contains person properties and attributes (e.g., name, email etc.).
Admin	Class inherited from person class, represent admin of the system and admin's attributes (e.g., phone number, admin code).
User	Class inherited from person class, represent user account of the system and user's attributes (e.g., emotion, voice etc.).
QsAnswer	Include all questions to be answered and their attributes (e.g., type, respond etc.).
Voice	Include voice recorded by the user and its attributes (e.g., voice, duration etc.).
Doctor	Include doctors contracting with system (e.g., degree, phone number).
VoiceEmotion	Include user feeling according to voice recorded (e.g., voice and emotion).
facialExpression	Include user feeling according to video recorded (e.g., video and emotion).

3.2.3 Sequence Diagrams

- Figure (3.4) below illustrate sequence diagram of voice emotion detection main function:

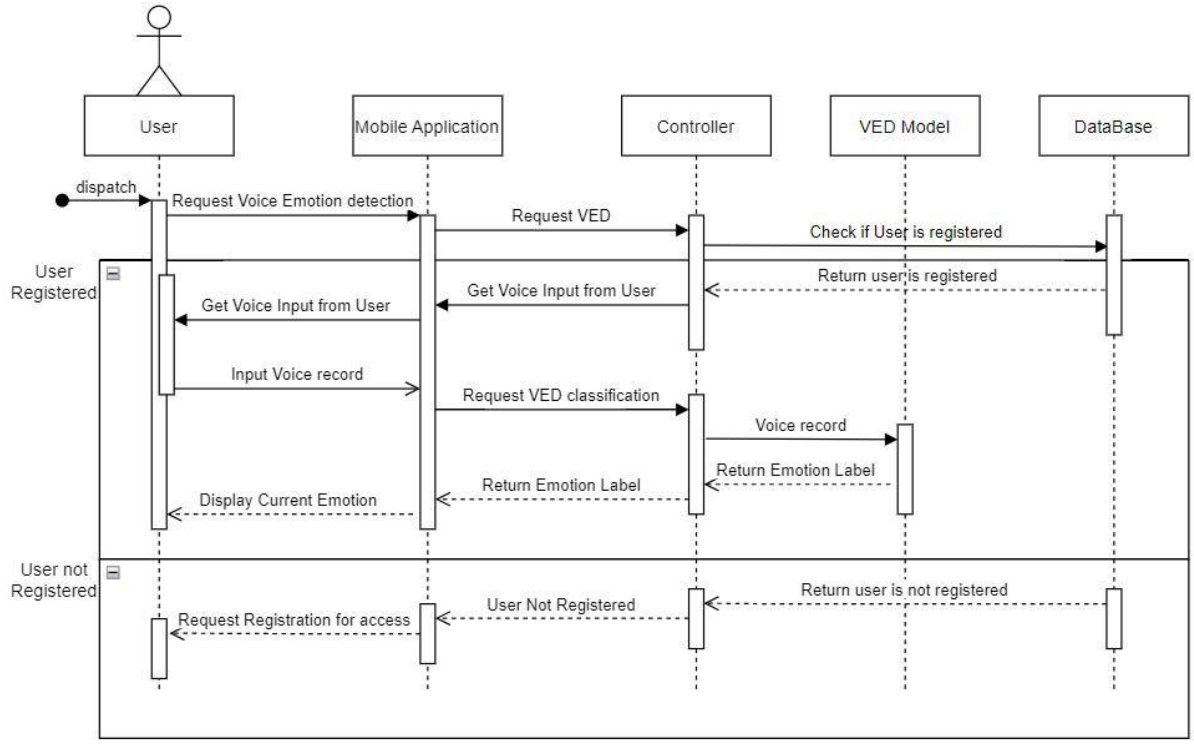


Figure3.4: VED sequence Diagram

Chapter 4

Implementation

4.1 Questionnaire

18 questions depend on the Psychological Wellbeing Scale, each question measure the state of the user by choosing different answers, The answer will as follow: There are seven levels to any question the weight of each question differs depends on the type of the question of positive type question with positive feedback in agreement, or negative type question that indicate negative mental issue in agreement, differentiation between the two types of questions in the way of assigning scores between Strongly Agree to Strongly Disagree.

4.1.1 Questions Construction:

Developed by psychologist Carol D. Ryff, the 18-item Psychological Wellbeing (PWB) Scale measures six aspects of wellbeing and happiness: autonomy, environmental mastery, personal growth, positive relations with others, purpose in life, and self-acceptance. Derived from the 42-item PWB to catch major aspects from the 42 in summarized 18 questions only.

4.1.2 Questions Data:

1. “I like most parts of my personality.”
2. “When I look at the story of my life, I am pleased with how things have turned out so far.”
3. “Some people wander aimlessly through life, but I am not one of them.”
4. “The demands of everyday life often get me down.”
5. “In many ways I feel disappointed about my achievements in life.”
6. “Maintaining close relationships has been difficult and frustrating for me.”
7. “I live life one day at a time and don’t really think about the future.”
8. “In general, I feel I am in charge of the situation in which I live.”
9. “I am good at managing the responsibilities of daily life.”
10. “I sometimes feel as if I’ve done all there is to do in life.”

11. “For me, life has been a continuous process of learning, changing, and growth.”
12. “I think it is important to have new experiences that challenge how I think about myself and the world.”
13. “People would describe me as a giving person, willing to share my time with others.”
14. “I gave up trying to make big improvements or changes in my life a long time ago”
15. “I tend to be influenced by people with strong opinions”
16. “I have not experienced many warm and trusting relationships with others.”
17. “I have confidence in my own opinions, even if they are different from the way most other people think.”
18. “I judge myself by what I think is important, not by the values of what others think is important.”

4.1.3 Scaling and measurement:

The PWB Scale has six subscales:

1 Autonomy (e.g., “I have confidence in my opinions, even if they are contrary to the general consensus”) Q15, Q17, Q18

2 Environmental Mastery (e.g., “In general, I feel I am in charge of the situation in which I live”) Q4, Q8, Q9

3 Personal Growth (e.g., “I think it is important to have new experiences that challenge how you think about yourself and the world”) Q11, Q12, Q14

4 Positive Relations with Others (e.g., “People would describe me as a giving person, willing to share my time with others”) Q6, Q13, Q16

5 Purpose in Life (e.g., “Some people wander aimlessly through life, but I am not one of them”) Q3, Q7, Q10

6 Self-acceptance (e.g., “When I look at the story of my life, I am pleased with how things have turned out”) Q1, Q2, Q5

4.1.4 Response Format

1 = strongly agree; 2 = somewhat agree; 3 = a little agree; 4 = neither agree nor disagree; 5 = a little disagree; 6 = somewhat disagree; 7 = strongly disagree

Q1, Q2, Q3, Q8, Q9, Q11, Q12, Q13, Q17, and Q18 are reverse scored

4.1.5 Results

The result of the users will be shown in a graph depends on their answers on the previous questions, The graph shows the user result in a blue column and below the graph an advice message will be shown,

- Firstly:
  If the result of all questions is less than or equal 42 the user score will be in the left column of the graph and the advice message will be "You Need to Go to Voice test and Image Test to get more accurate result"
- Secondly:
  If the result of all questions is from 43:83 the user score will be in the middle column of the graph and the advice message will be "You need to meet new people with positive energy and love for life, Participation in some social activities and charities And Reading especially books that give life joy."
- Thirdly:
  If the result of all questions is greater than 84 the user score will be in the right column of the graph and the advice message will be "You are mentally stable"

4.2 Voice Emotion Detection (VED)

Voice Emotion Detection, abbreviated as VED, is the act of attempting to recognize human emotion and affective states from speech. This is capitalizing on the fact that voice often reflects underlying emotion through tone and pitch. This is also the phenomenon that animals like dogs and horses employ to be able to understand human emotion.

1. Emotion recognition is the part of speech recognition which is gaining more popularity and need for it increases enormously. Although there are methods to recognize emotion using machine learning techniques, this project attempts to use deep learning to recognize the emotions from data.
2. VED (Voice Emotion Detection) is used in call center for classifying calls according to emotions and can be used as the performance parameter for conversational analysis thus identifying the unsatisfied customer, customer satisfaction and so on... for helping companies improving their services
3. It can also be used in-car board system based on information of the mental state of the driver can be provided to the system to initiate his/her safety preventing accidents to happen

4.2.1 Datasets used in project:

1- Crowd-sourced Emotional Mutimodal Actors Dataset (Crema-D)
2- Ryerson Audio-Visual Database of Emotional Speech and Song (Ravdess)
3- Surrey Audio-Visual Expressed Emotion (Savee)
4- Toronto emotional speech set (Tess)

4.2.2 Data Collection

Working with four different datasets: will be creating a data frame storing all emotions of the data in data frame with their paths

Will use this data frame to extract features for our model training.

1. Ravdess Dataframe

Here are the filename identifiers as per the official RAVDESS website:

- Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
- Vocal channel (01 = speech, 02 = song).
- Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
- Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion.
- Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
- Repetition (01 = 1st repetition, 02 = 2nd repetition).
- Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).

So, here's an example of an audio filename. 02-01-06-01-02-01-12.mp4 This means the meta data for the audio file is:

- Video-only (02)
- Speech (01)
- Fearful (06)
- Normal intensity (01)
- Statement "dogs" (02)
- 1st Repetition (01)
- 12th Actor (12) - Female (as the actor ID number is even)

2. Crema DataFrame

Crema Dataframe is simply organized to ‘SAD’ (sad), ‘ANG’ (angry), ‘DIS’ (disgust), ‘FEA’ (fear), ‘HAP’ (happy), ‘NEU’ (neutral).

3. TESS dataset

TESS dataset is simply organized to sad, angry, disgust, fear, happy, neutral, pleasant_surprised changing into surprise.

4. CREMA-D dataset

The audio files in this dataset are named in such a way that the prefix letters describe the emotion classes as follows:

- 'a' = 'anger'
- 'd' = 'disgust'
- 'f' = 'fear'
- 'h' = 'happiness'
- 'n' = 'neutral'
- 'sa' = 'sadness'
- 'su' = 'surprise'

4.2.3 Data Visualization and Exploration

In Figure (4.1) below the Emotion count after data preparation from 4 datasets.

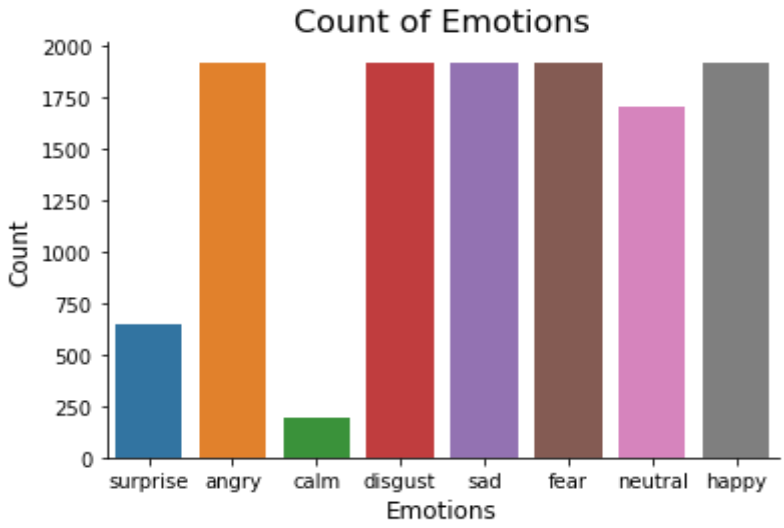


Figure 4.1: Count of each emotion

Plot waveplots and spectrograms for audio signals to examine the frequency that would work on

- Waveplots - Waveplots let us know the loudness of the audio at a given time.
- Spectrograms - A spectrogram is a visual representation of the spectrum of frequencies of sound or other signals as they vary with time. It's a representation of frequencies changing with respect to time for given audio/music signals.

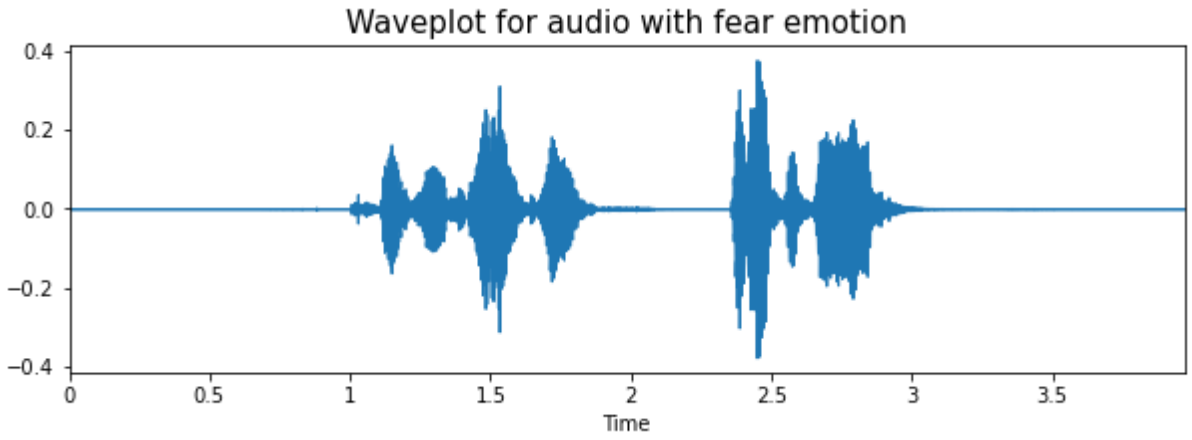


Figure4.2 : Wave plot for audio with fear emotion

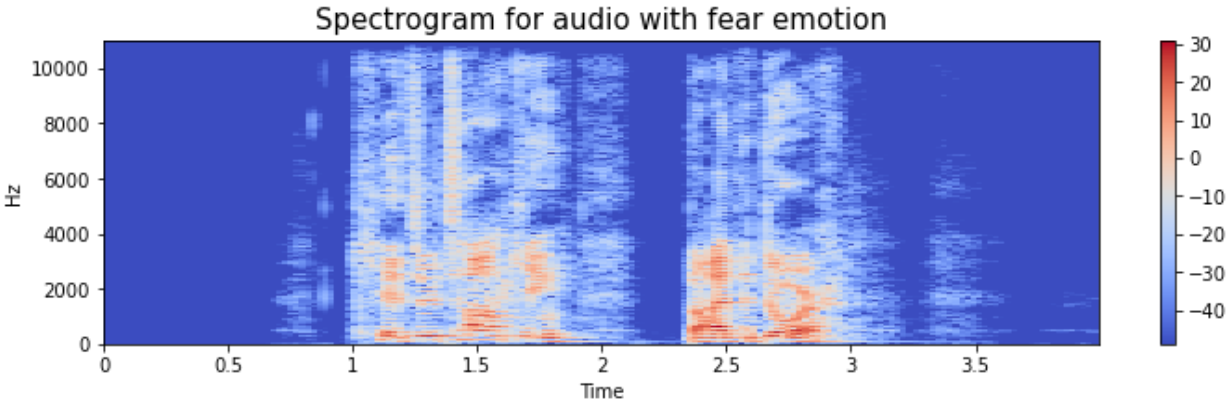


Figure4.3: Spectrogram for audio with fear emotion

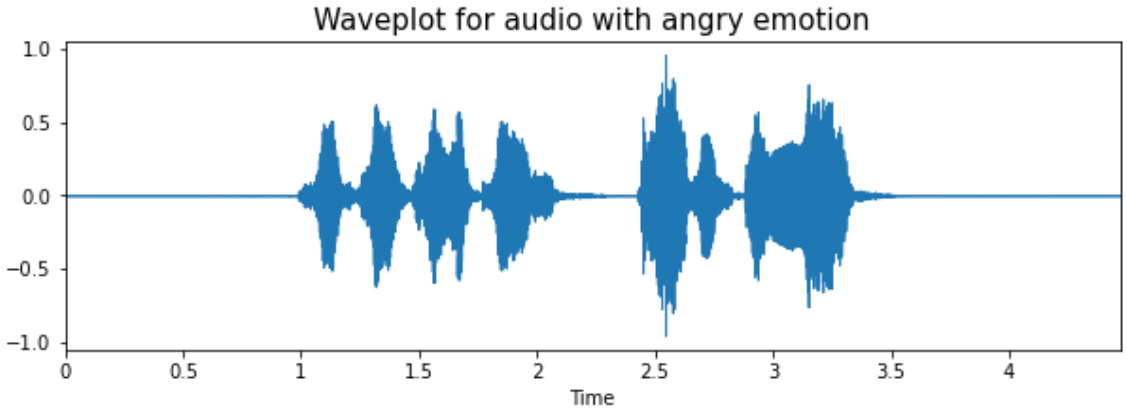


Figure4.4: Wave plot for audio with angry emotion

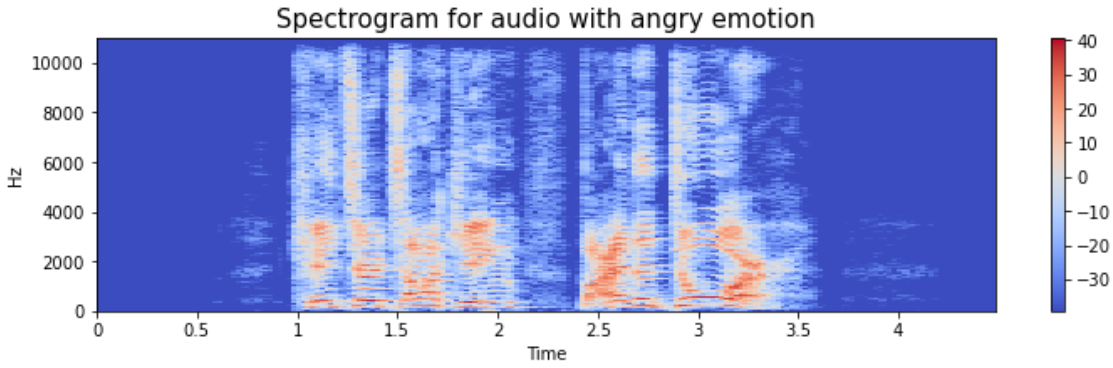


Figure4.5: Spectrogram for audio with angry emotion

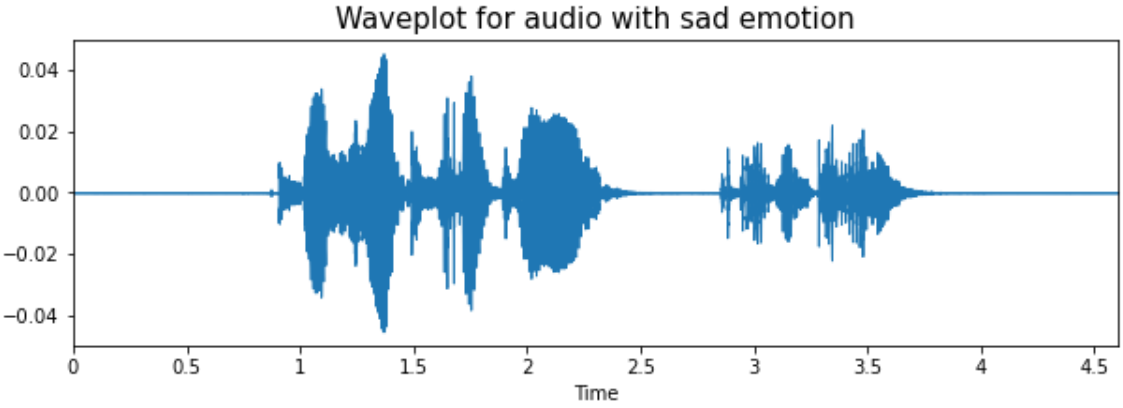


Figure4.6: Wave plot for audio with sad emotion

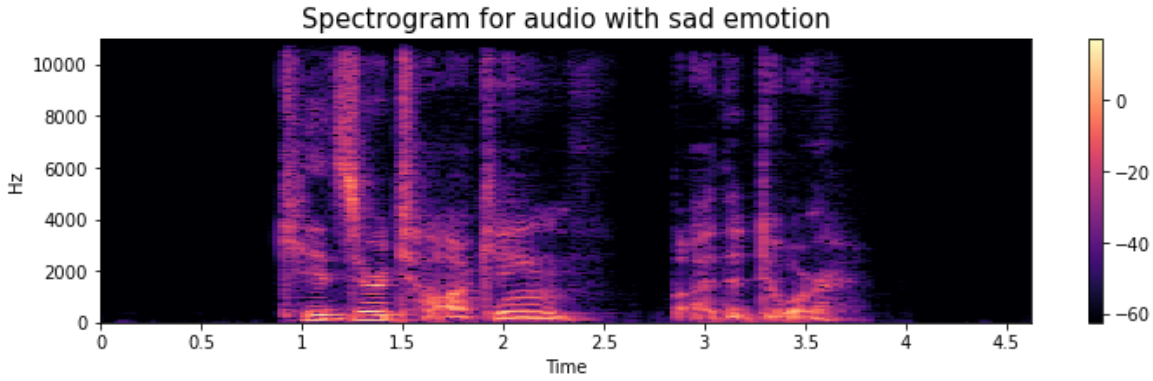


Figure4.7: Spectrogram for audio with sad emotion

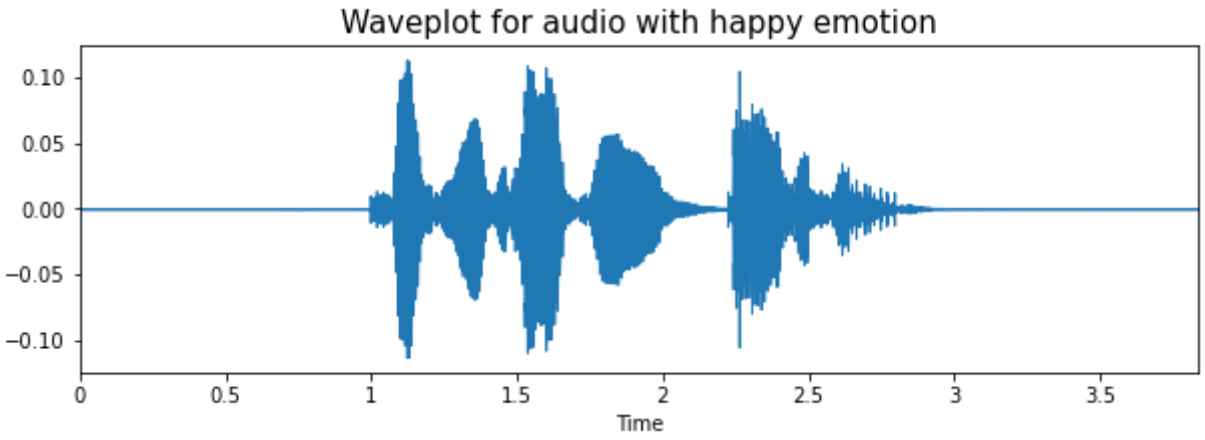


Figure4.8: Wave plot for audio with happy emotion

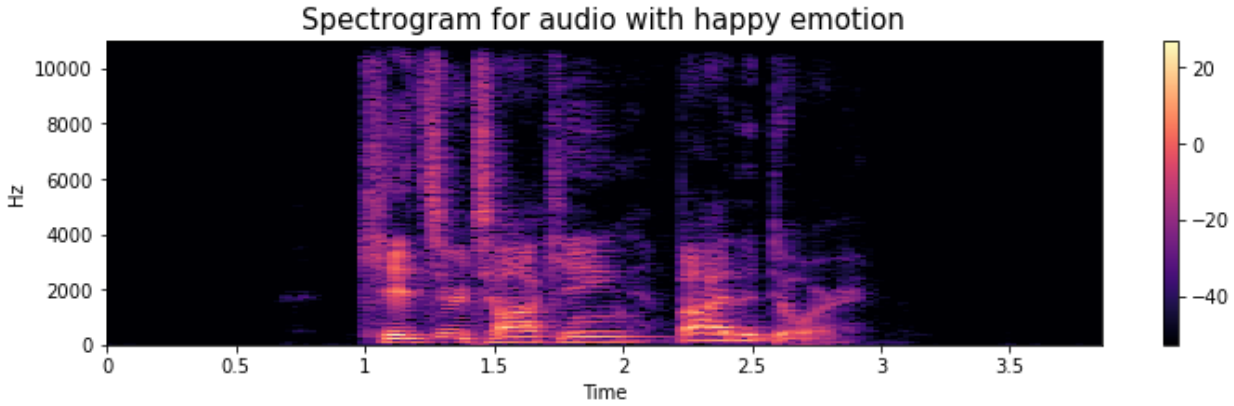


Figure4.9: Spectrogram for audio with happy emotion

4.2.4 Data Augmentation

- Data augmentation is the process by which create new synthetic data samples by adding small perturbations on our initial training set.
- To generate syntactic data for audio, applying noise injection, shifting time, changing pitch and speed.
- The objective is to make our model invariant to those perturbations and enhance its ability to generalize.
- For this to work adding the perturbations must conserve the same label as the original training sample.
- In images data augmentation can be performed by shifting the image, zooming, rotating ...

First, let's check which augmentation techniques works better for our dataset.

Next Figure (4.10) of Simple Audio

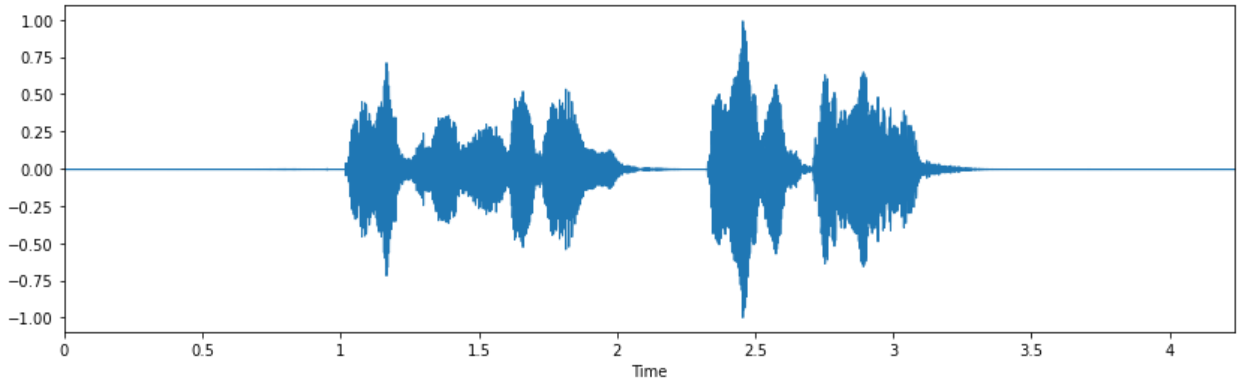


Figure 4.10: Simple Audio

Next Figure (4.11) Noise Injection into Audio

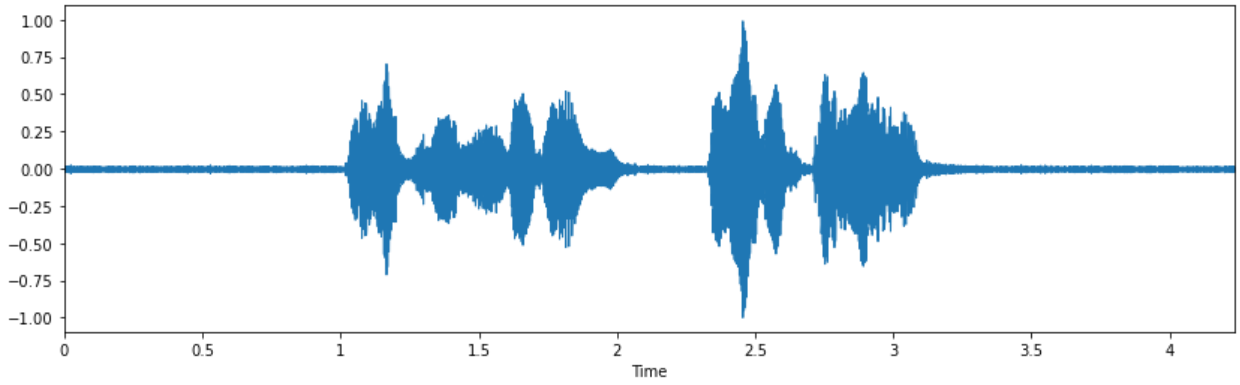


Figure4.11: Noise injection into audio

Noise injection is a very good augmentation technique because of which we can assure our training model is not overfitted.

Next Figure (4.12) Stretching of Audio

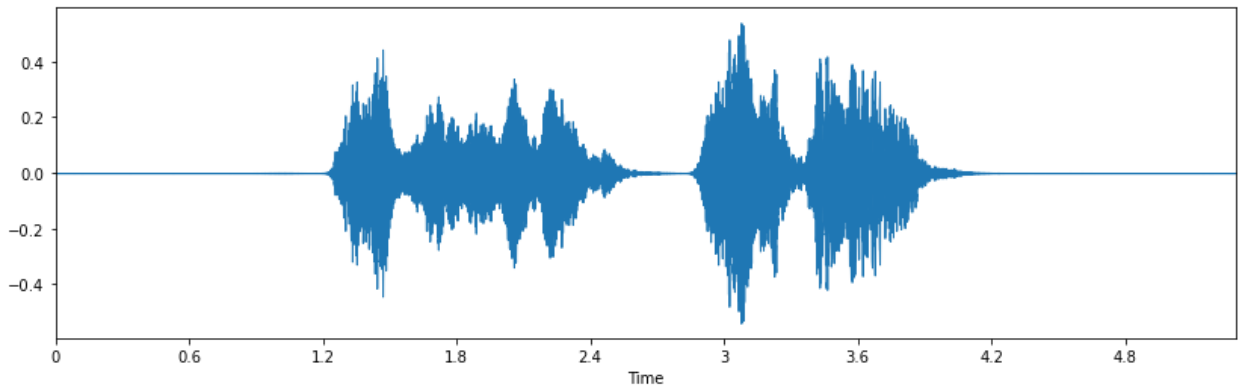


Figure4.12: Stretching of Audio

Next Figure (4.13) Shifting of Audio

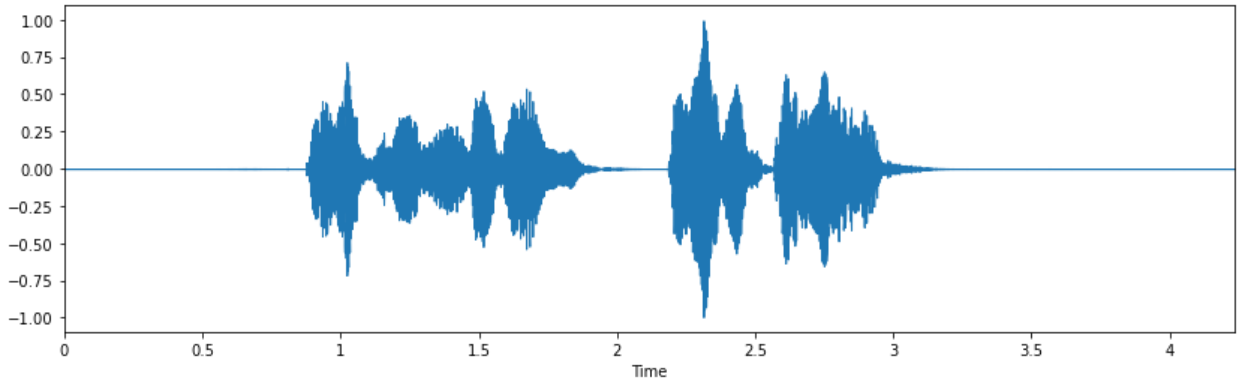


Figure4.13: Shifting of Audio

Next Figure (4.14) changing Pitch of Audio

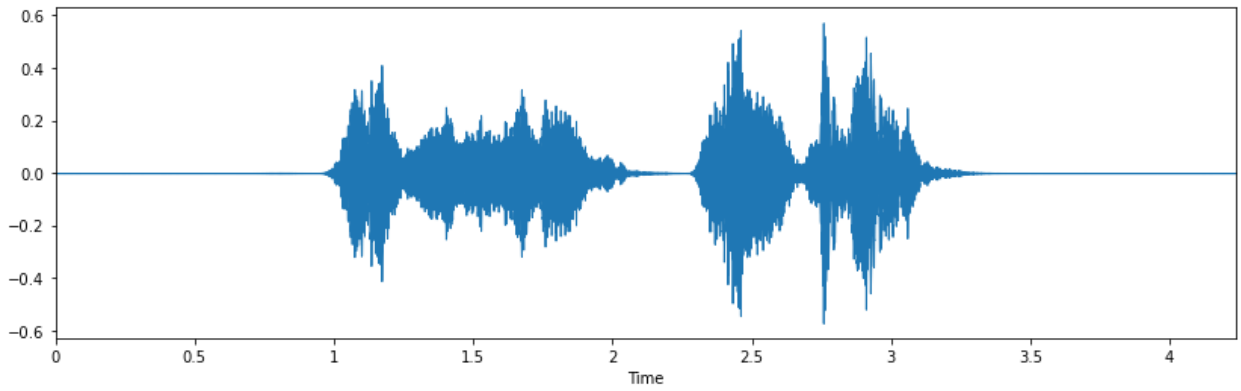


Figure4.14: changing Pitch of Audio

From the above types of augmentation techniques noise, stretching (ie. changing speed) and some pitching is used. Not using shifting as it does not affect the frequency thus result not affected in detection.

Thus, Data size from 12,162 after Augmentation by applying Noise, Stretching with Pitching into 36,486 audio files

4.2.5 Feature Extraction

Analyzing and finding relations between different things. As already known that the data provided of audio cannot be understood by the models directly so its needed to be converted into an understandable format for which feature extraction is used.

The audio signal is a three-dimensional signal in which three axes represent time, amplitude, and frequency.

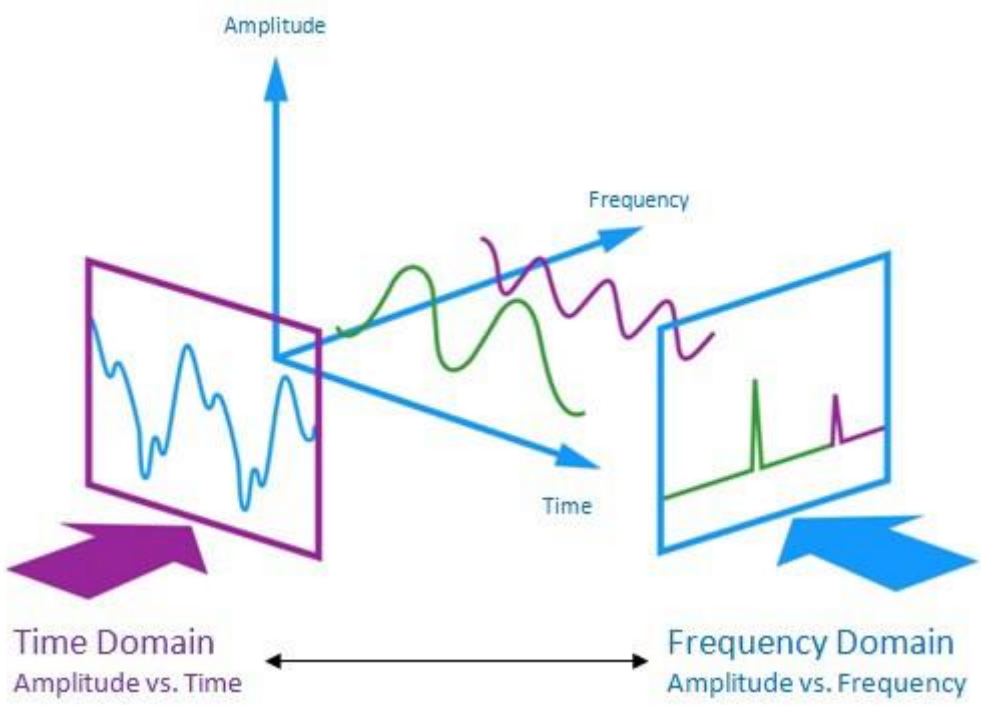


Figure4.15: Feature Extraction

With the help of the sample rate and the sample data, one can perform several transformations on it to extract valuable features out of it.

1. Zero Crossing Rate: The rate of sign-changes of the signal during the duration of a particular frame.
2. Energy: The sum of squares of the signal values, normalized by the respective frame length.
3. Entropy of Energy: The entropy of sub-frames' normalized energies. It can be interpreted as a measure of abrupt changes.
4. Spectral Centroid: The center of gravity of the spectrum.
5. Spectral Spread: The second central moment of the spectrum.

6. Spectral Entropy: Entropy of the normalized spectral energies for a set of sub-frames.
7. Spectral Flux: The squared difference between the normalized magnitudes of the spectra of the two successive frames.
8. Spectral Rolloff: The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
9. MFCCs Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale.
10. Chroma Vector: A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).
11. Chroma Deviation: The standard deviation of the 12 chroma coefficients.

In this project: only extracting 5 features:

- Zero Crossing Rate
- Chroma_stft
- MFCC
- RMS (root mean square) value
- MelSpectogram to train our model

Scaling data using StandardScaler and split our data for training and testing with ratio 3:1

Making Data compatible to Model: by Expanding extra dimension on axis 2

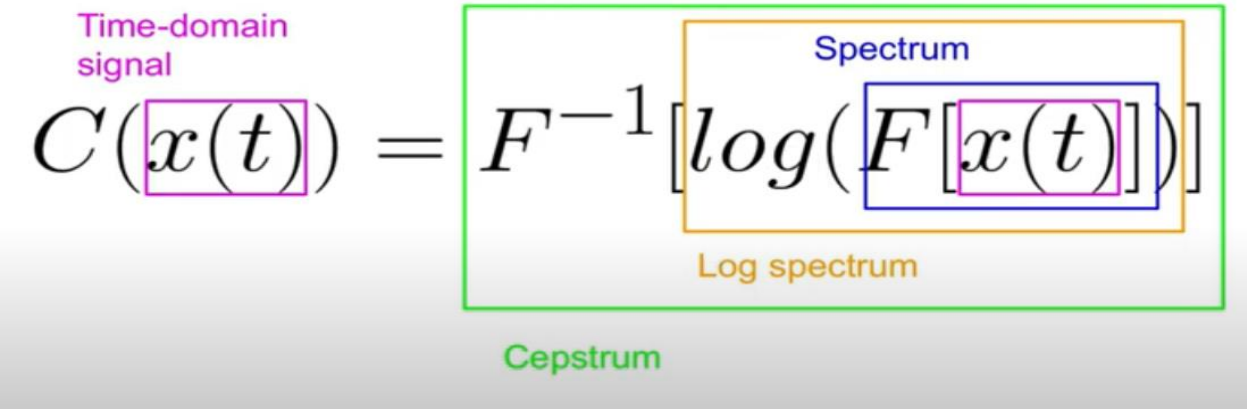


Figure 4.16 MFCC equation

4.2.6 Modelling

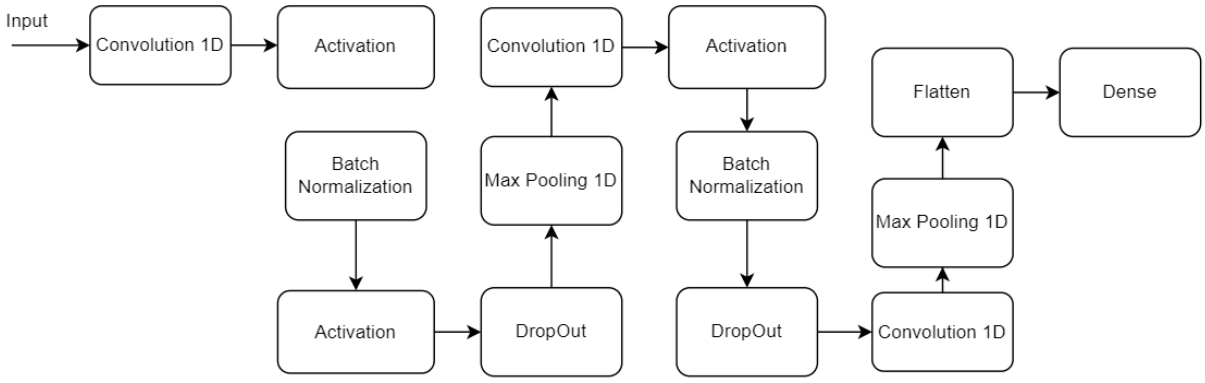


Figure4.17: VED CNN Model layers

Create a network that takes as input 162-by-1 images and the corresponding labels and outputs a scalar prediction score using a series of convolution layers with batch normalization and ReLU layers. Add noise to the input images using dropout.

- For the dropout layer, specify a dropout probability of 0.25.
- For the convolution layers, specify 256 filters with decreasing number of filters for each layer and kernel size of 8.
- Also specify pooling and batching with normalization for values.
- Flat values for easier interaction and Dense to desired value
- Work on 428 Epochs with Reduce Learning Rate on Plateu
- Faster learning in early epochs to exit after non-learning for specific epochs Min Learning Rate 0.0000001

4.2.7 Evaluation

Accuracy of our model on test data: 73.43783974647522 %

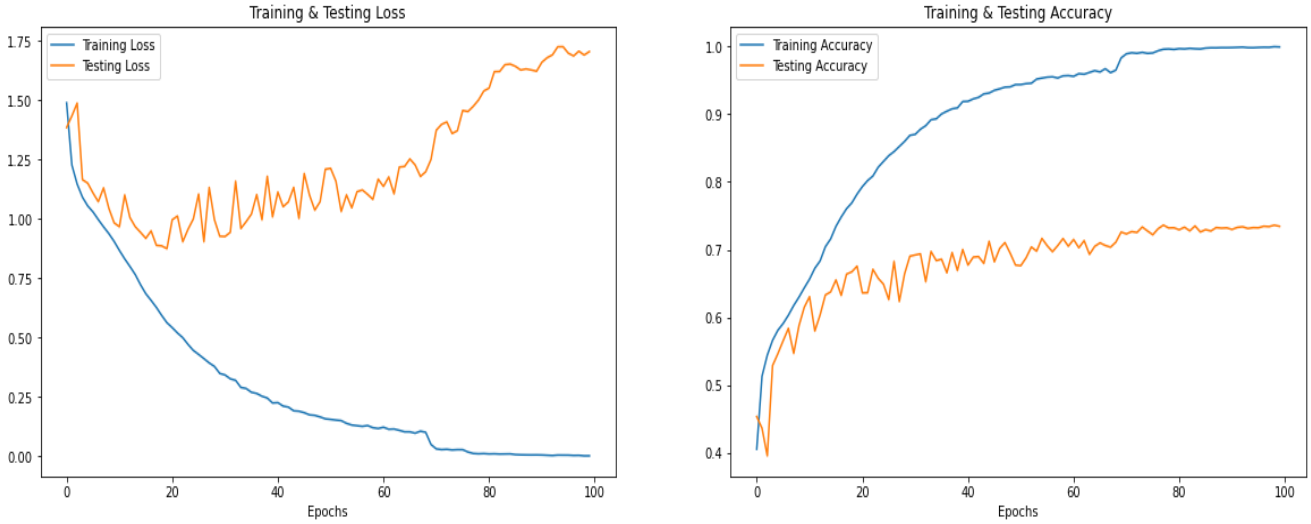


Figure4.18: Train/Test Loss&Accuracy

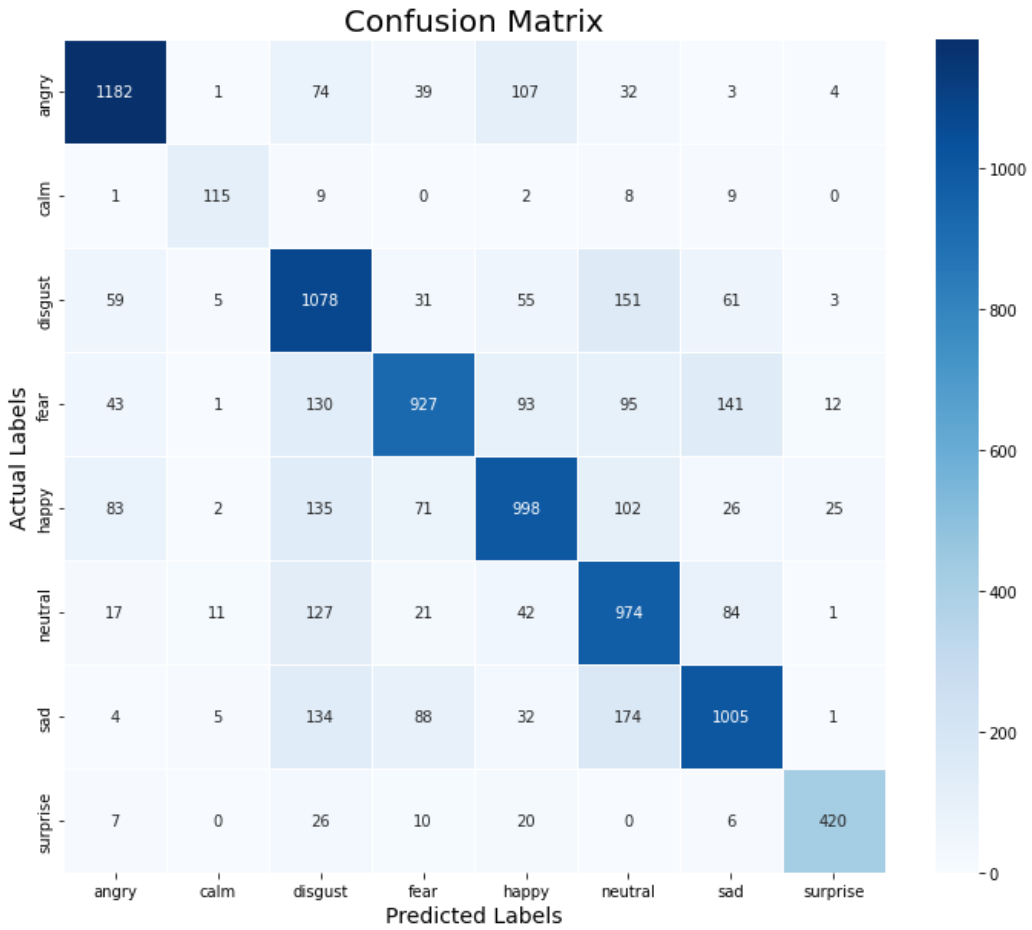


Figure 4.19 Confusion Matrix of Model

4.3 Facial Emotion Detection (FED)

Facial emotion recognition is the task of classifying the expressions on face images into various categories such as anger, fear, surprise, sadness, happiness and so on

4.3.1 Datasets and Description

FER2013 (Facial Expression Recognition 2013 Dataset)

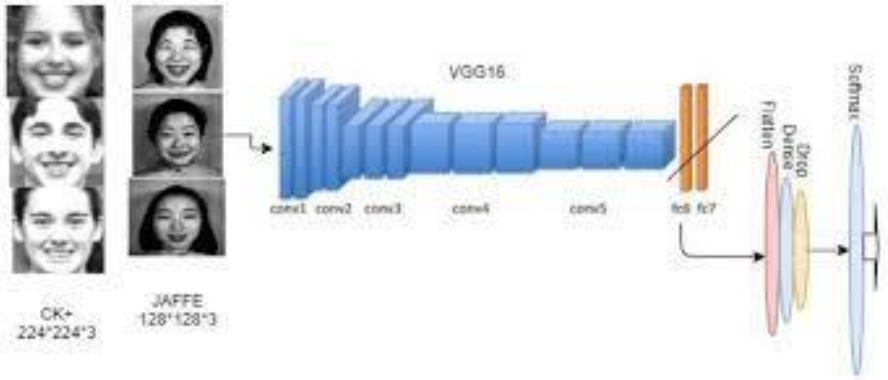


Figure:4.20 Facial Expression Recognition

Fer2013 contains approximately 30,000 facial RGB images of different expressions with size restricted to 48x48 pixel grayscale images of faces. The faces have been automatically registered so that the face is centred and occupies about the same amount of space in each image.

The task is to categorize each face based on the emotion shown in the facial expression into one of seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral). The training set consists of 28,709 examples and the public test set consists of 3,589 examples.

The Disgust expression has the minimal number of images – 600, while other labels have nearly 5,000 samples each.

4.3.2 Data Preparation and Collection

'Anger' (0, 2652), 'Disgust' (1, 2694), 'Fear' (2, 2677), 'Happy' (3, 2668), 'Neutral' (4, 2701), 'Sad' (5, 2554), 'Surprise' (6, 2657)

Figure (4.21) below of images data according to the labeled emotion:

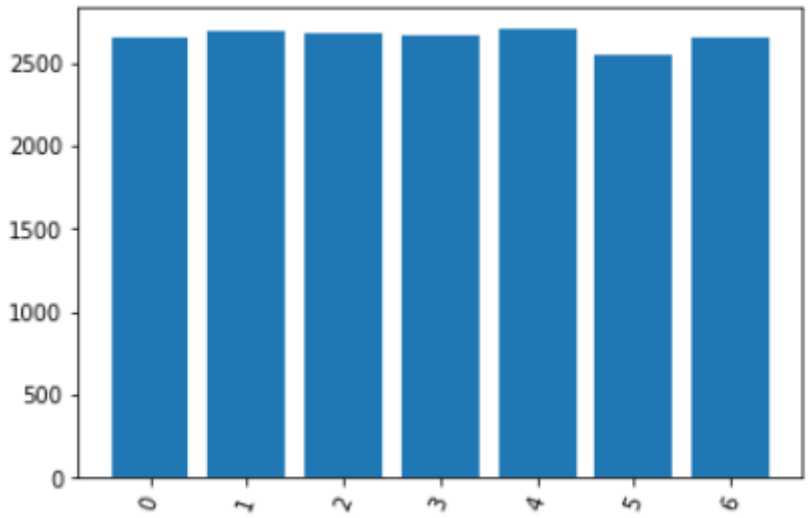


Figure4.21 : images data according to the labeled emotion

4.3.3 Modelling

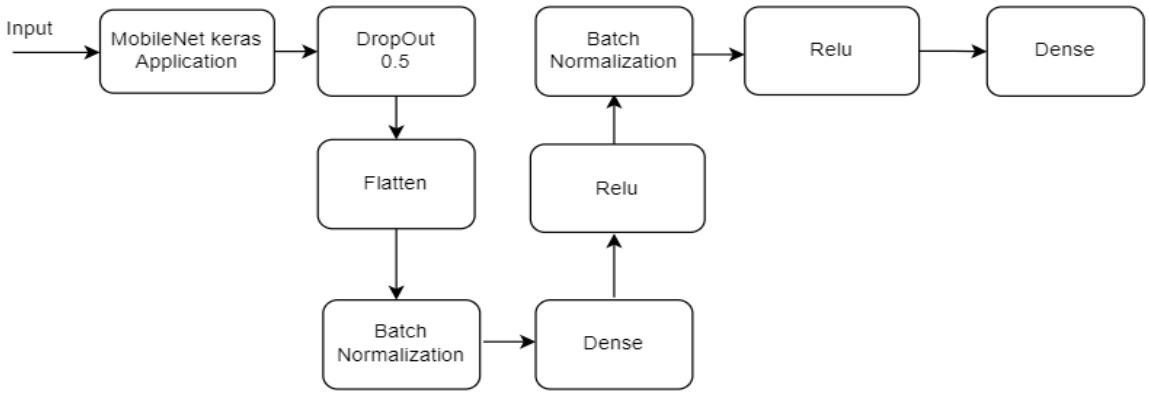
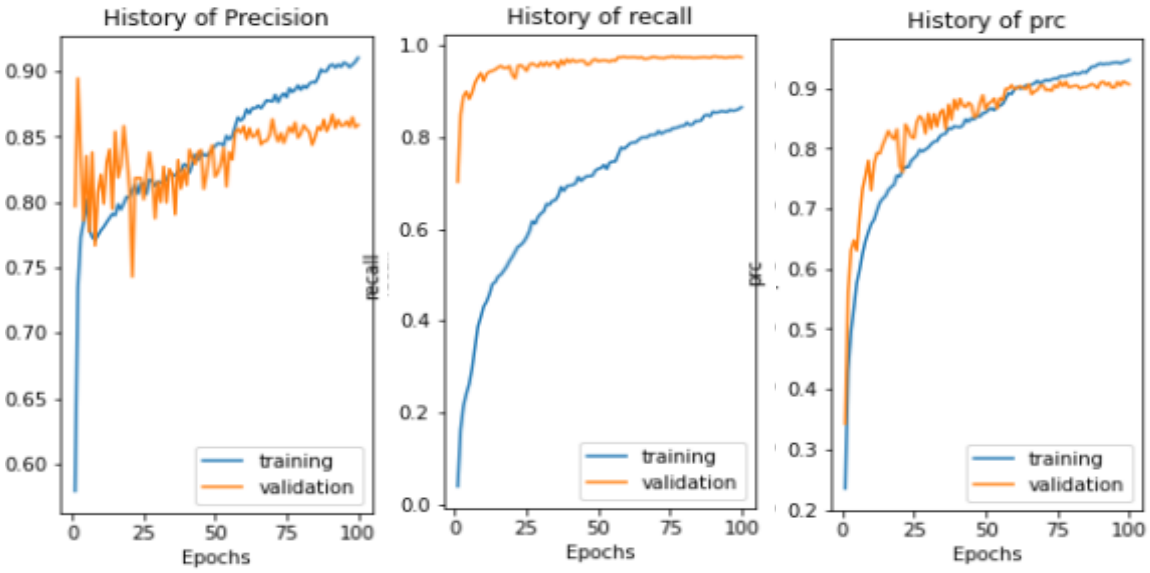


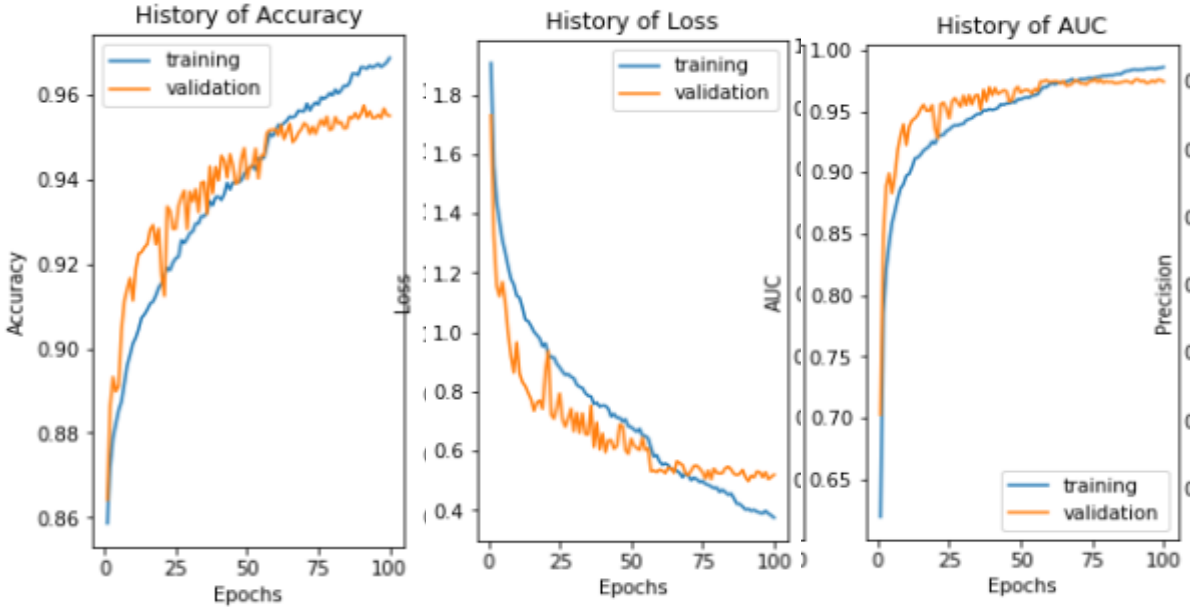
Figure 4.22: FED CNN model layers

Create a efficient Convolutional Neural Network for Mobile Vision applications (streamlined architecture that uses depth-wise separable convolutions to build light weight deep neural networks) that takes as input 48-by-48-by-3 images and the corresponding labels and outputs a scalar prediction score using a series of convolution layers with batch normalization and ReLU layers.

- For the dropout layer, specify a dropout probability of 0.5.
- Also specify pooling and batching with normalization for values.
- Flat values for easier interaction and Dense to desired value
- Work on 100 Epochs

4.3.4 Evaluation





CNN

CNN is a neural network: an algorithm used to recognize patterns in data. Neural Networks in general are composed of a collection of neurons that are organized in layers, each with its own learnable weights and biases, it has become an important tool for object recognition, action classification, and image captioning.
CNN consists of input layer and an output layer, as well as multiple hidden layers.

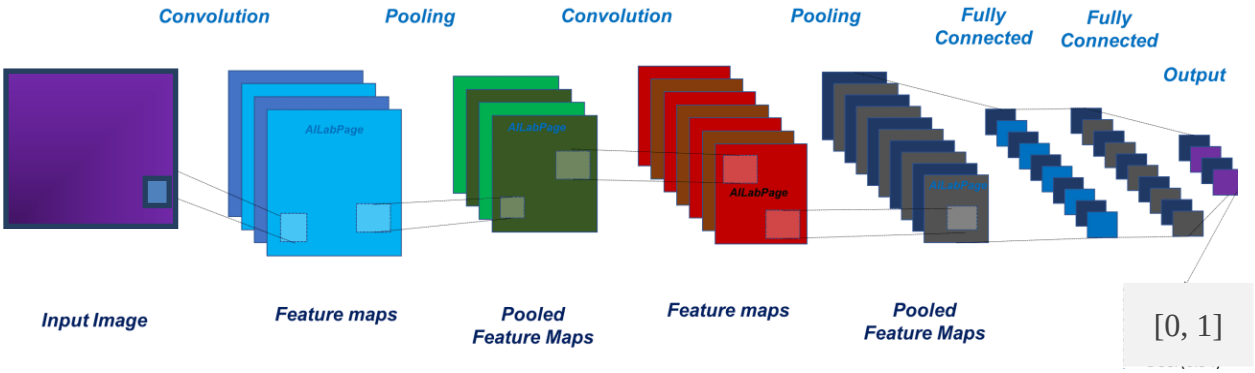


Figure4.23: CNN basic architecture

Input layer: represents the input image into the CNN. Because inputs are RGB images, the input layer has three channels, corresponding to the red, green, and blue channels, respectively.

The Multiple hidden layers: filter (convolve) the inputs to get useful information, it consists of convolution layers, pooling layers, fully connected layers and normalization layer.

Convolutional Layers: The convolutional layers are the foundation of CNN, as they contain the learned kernels (weights), which extract features that distinguish different images from one another, this is what we want for classification. As you interact with the convolutional layer, you will notice links between the previous layers and the convolutional layers. Each link represents a unique weight used for the convolution operation to produce the current convolutional neuron's output or activation map.
The convolutional neuron performs an element-wise dot product with a unique kernel and the previous layer's corresponding neuron output. This will yield as many intermediate results as there are unique kernels. The convolutional neuron is the result of all the intermediate results summed together with the learned bias.

Convolution operation: is a mathematical operation derived from the domain of signal processing. It is used to get an output given the model and the input. The model is represented as a transfer function. The input convolved with the transfer function results in the output.

1. **Padding** is often necessary when the kernel extends beyond the activation map. Padding conserves data at the borders of activation maps, which leads to better performance, and it can help preserve the input's spatial size, which allows an architecture designer to build higher performing networks. There exist many padding techniques, but the most commonly used approach is zero-padding because of its performance, simplicity, and computational efficiency. The technique involves adding zeros symmetrically around the edges of an input. This approach is adopted by many high-performing CNNs.

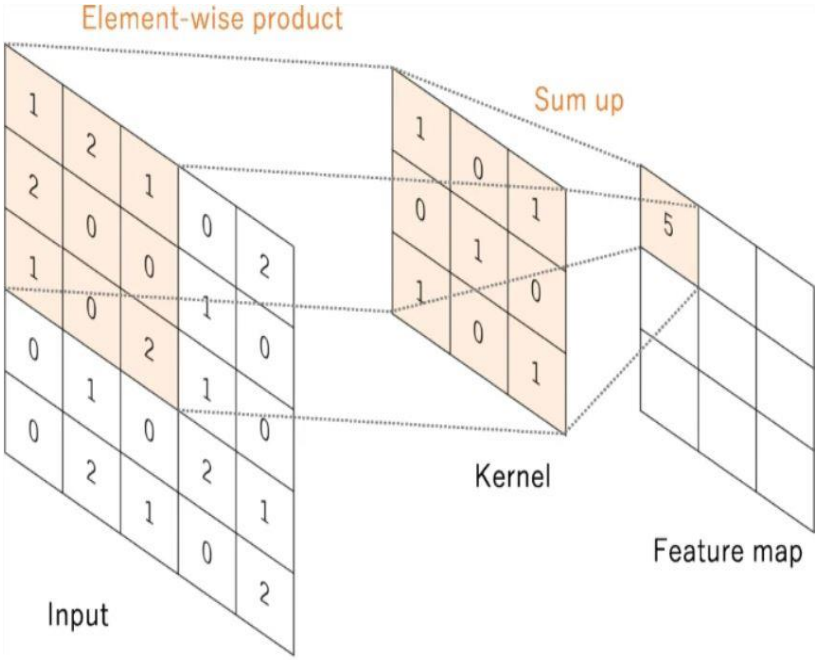


Figure4.24: convolution operation

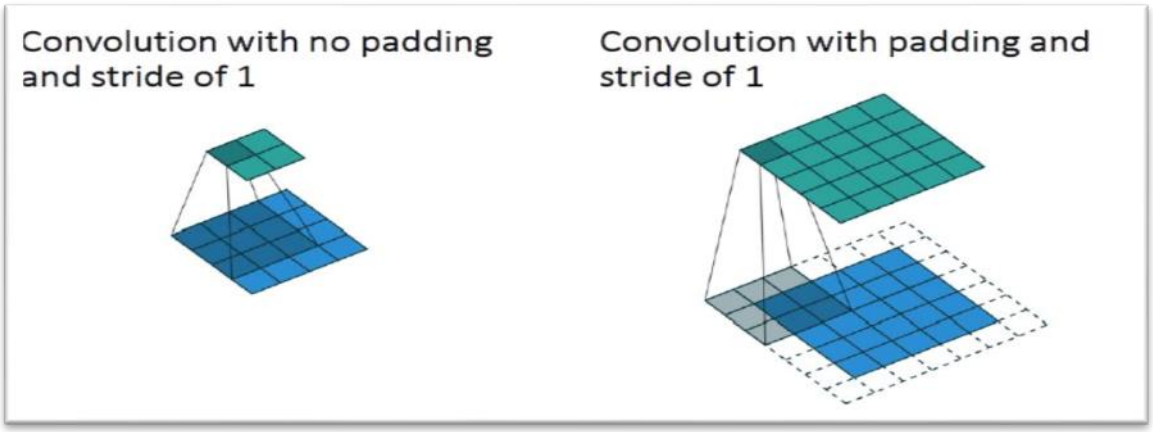


Figure4.25: padding

Output size for 2D:

$$outputSize_W = (W - F + 2P)/S + 1$$
$$outputSize_H = (H - F + 2P)/S + 1$$

2. **Kernel size**, often also referred to as filter size, refers to the dimensions of the sliding window over the input. Choosing this hyper parameter has a massive impact on the image classification task. For example, small kernel sizes are able to extract a much larger amount of information containing highly local features from the input. As you can see on the visualization above, a smaller kernel size also leads to a smaller reduction in layer dimensions, which allows for a deeper architecture. Conversely, a large kernel size extracts less information, which leads to a faster reduction in layer dimensions, often leading to worse performance. Large kernels are better suited to extract features that are larger. At the end of the day, choosing an appropriate kernel size will be dependent on your task and dataset, but generally, smaller kernel sizes lead to better performance for the image classification task because an architecture designer is able to stack more and more layers together to learn more and more complex features!

3. **Stride** indicates how many pixels the kernel should be shifted over at a time.

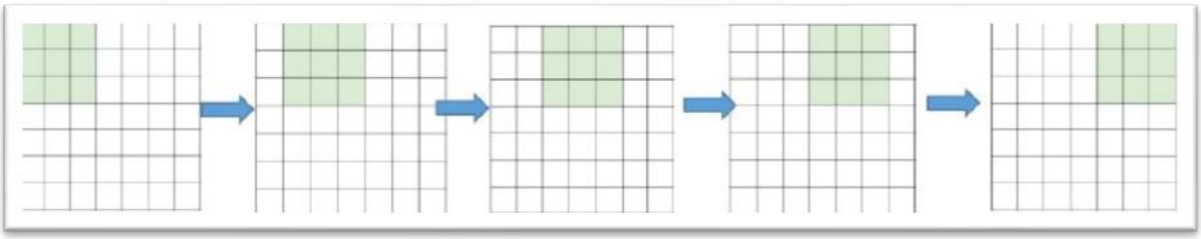


Figure4.26: stride =1

Activation Functions:

Convolution is a linear operation; therefore, we need a non-linearity, and otherwise two convolution layers would be no more powerful than one.

-ReLU

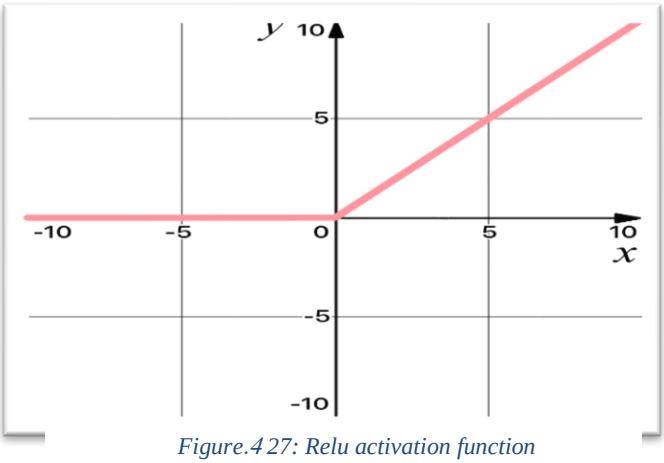


Figure.4 27: Relu activation function

Neural networks are extremely prevalent in modern technology—because they are so accurate! The highest performing CNNs today consist of an absurd amount of layers, which are able to learn more and more features. ReLU applies much-needed non-linearity to the model. Non-linearity is necessary to produce non-linear decision boundaries so that the output cannot be written as a linear combination of the inputs. If a non-linear activation function was not present, deep CNN architectures would devolve into a single, equivalent convolutional layer, which would not perform nearly as well. The ReLU activation function is specifically used as a non-linear activation function, as opposed to other non-linear functions such as *Sigmoid* because it has been empirically observed that CNNs using ReLU are faster to train than their counterparts.

-Sigmoid

It is one of the most widely used non-linear activation function. Sigmoid transforms the values between the range 0 and 1. Here is the mathematical expression for sigmoid: $f(x) = 1/ (1+\exp(-x))$

-Tanh:

The tanh function is very similar to the sigmoid function. The only difference is that it is symmetric around the origin. The range of values in this case is from -1 to 1. Thus the inputs to the next layers will not always be of the same sign. The tanh function is defined as:

$$Tanh(x) = 2sigmoid(2x) - 1$$

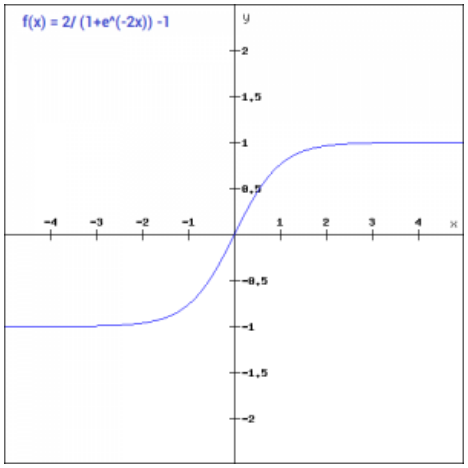


Figure .428 : tanh activation function

Pooling Layers

There are many types of pooling layers in different CNN architectures, but they all have the purpose of gradually decreasing the spatial extent of the network, which reduces the parameters and overall computation of the network.

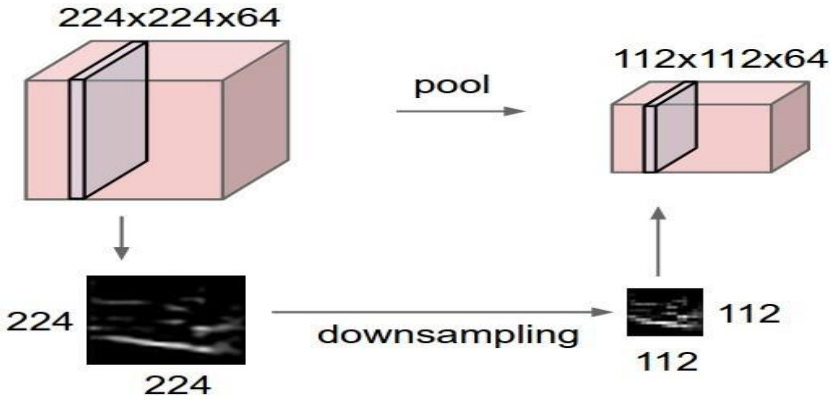


Figure4.29: pooling

Pooling layer makes the input representation (feature dimension) smaller and more manageable, reduce the number of parameters and computations in the network, therefore, controlling Overfitting, makes the network invariant to small transformations, distortions and translations in the input image (a small distortion in input will not change the output of pooling – since we take the maximum / average value in a local neighborhood).

*Overfitting is to train training set too much and make the gap between test error and validation error very large.

Normalization

Normalization is a pre-processing technique used to standardize data. In other words, having different sources of data inside the same range. Not normalizing the data before training can cause problems in our network, making it harder to train and decrease its learning speed.

There are two main methods to normalize our data. The most straightforward method is to scale it to a range from 0 to 1:

$$\mathbf{Xnormalized = (X-M) / (Xmax - Xmin)}$$

X the data point to normalize, M the mean of the data set, Xmax the highest value, and Xmin the lowest value. This technique is generally used in the inputs of the data. The non-normalized data points with wide ranges can cause instability in Neural Networks. The relatively large inputs can cascade down to the layers, causing problems such as exploding gradients.

The other technique used to normalize data is forcing the data points to have a mean of 0 and a standard deviation of 1, using the following formula:

$$\mathbf{Xnormalized = (X-M) / (S)}$$

Being X the data point to normalize, M the mean of the data set, and S the standard deviation of the data set. Now, each data point mimics a standard normal distribution. Having all the features on this scale, none of them will have a bias, and therefore, models will learn better.

Batch Normalization:

Batch Norm is a normalization technique done between the layers of a Neural Network instead of in the raw data. It is done along mini batches instead of the full data set. It serves to speed up training and use higher learning rates, making learning easier.

Fully Connected Layer

The output from the convolutional and pooling layers represents high-level features of the input image. The purpose of the Fully Connected layer is to use these features for classifying the input image into various classes based on the training dataset.

The Fully Connected layer is a traditional Multilayer Perceptron MLP.it is called the “output layer” and in classification settings it represents the class scores.

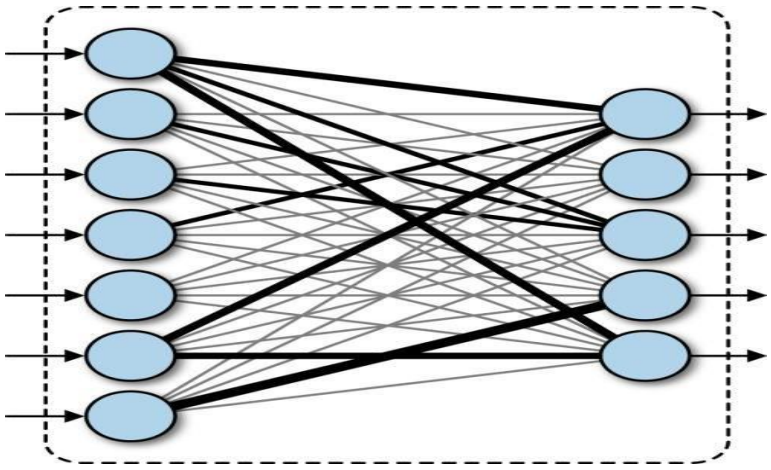


Figure.4.30: Fully connected layer

Downsampling

First step is to extract feature from image which is done a convolution network, as input a convolution layer network takes an image, size of filter window that move over input image to extract out features and the stride size.

Each convolution layer leads to extraction of progressively higher-level features, so the downsampling block consists of two convolution layers.

Residual Block:

To solve the problem of the vanishing/exploding gradient, this architecture introduced the concept called Residual Network. In this network we use a technique called ***skip connections***. The skip connection skips training from a few layers and connects directly to the output.

The approach behind this network is instead of layers learn the underlying mapping, we allow network to fit the residual mapping. So, instead of say H(x), initial mapping, let the network fit,

F(x): = H(x) – x which gives *H(x): = F(x) + x*.

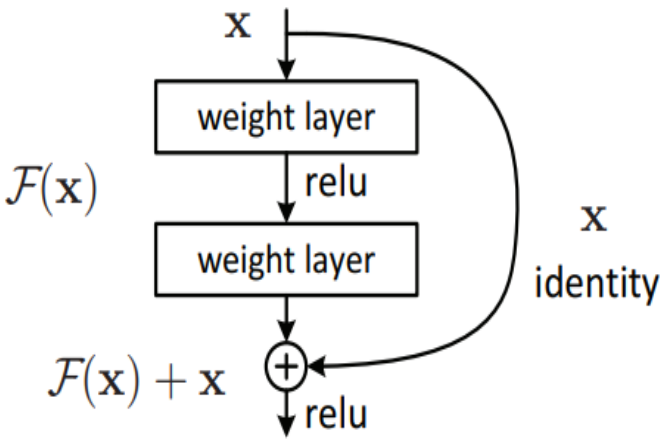


Figure 4.31: Resnet

The advantage of adding this type of skip connection is because if any layer hurt the performance of architecture, then it will be skipped by regularization. So, this results in training very deep neural network without the problems caused by vanishing/exploding gradient.

There is a similar approach called “highway networks”.

Resnet_block in the model is a neural network layer that consists of two convolution layers where a residue of input is added to the output. This is done to ensure properties of input of previous layers are available for

later layers as well so that their output does not deviate much from the original input, otherwise, the characteristics of original images will not be retained in the output and results will be very abrupt. One of the primary aims of the task is to retain the characteristic of the original input like the size and shape of the object, so residual networks are a great fit for these kinds of transformations. Resnet block can be summarized in the following image:

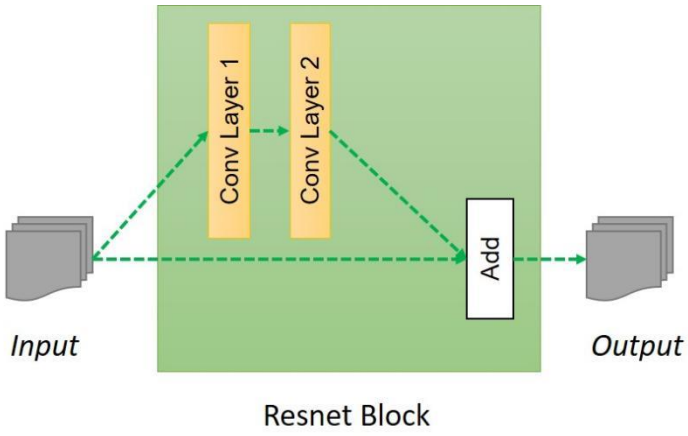


Figure 4.32: Resnet in CycleGAN

Upsampling

It is the exact opposite of downsampling, to build back the low-level features back from the features vector. This is done by applying a deconvolution (or transpose convolution) layer.

Optimization function

Adam: is a stochastic optimization method that has been proven useful for various types of network architectures. Adam uses an individual learning rate for each weight in a network and it is approximated using two values from the weights gradient and not only its average. The step size is decreasing based on the momentum. Adam is a popular algorithm in the field of deep learning because it achieves good results fast.

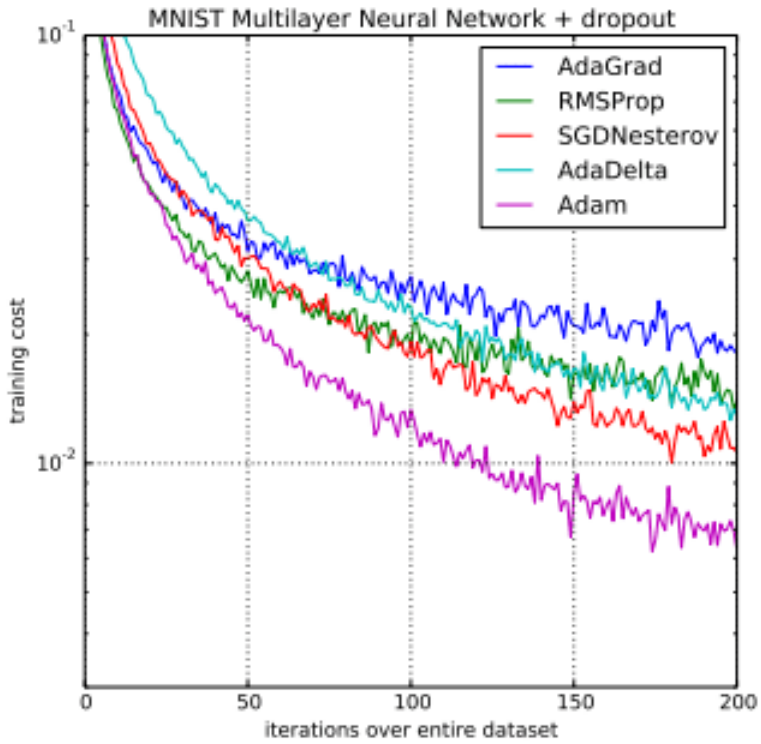


Figure 4.33: Adam optimizer

Chapter 5

User Manual

First when open application:

1-Login page :

There are 2 textboxes:

1- Enter mail
2- Enter password

Click on sign into login

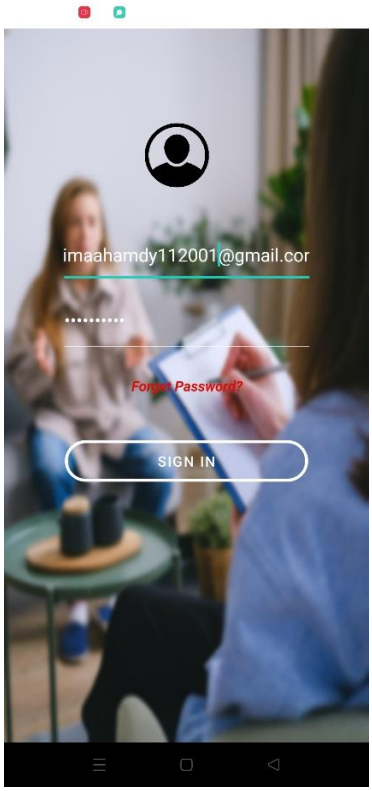


Figure 5.1: Login Page

2-Register page:

There are 5 textboxes:

1- Enter mail
2- Enter age
3- Enter password
4- Confirm password

Click on button sign up to save your data and regist

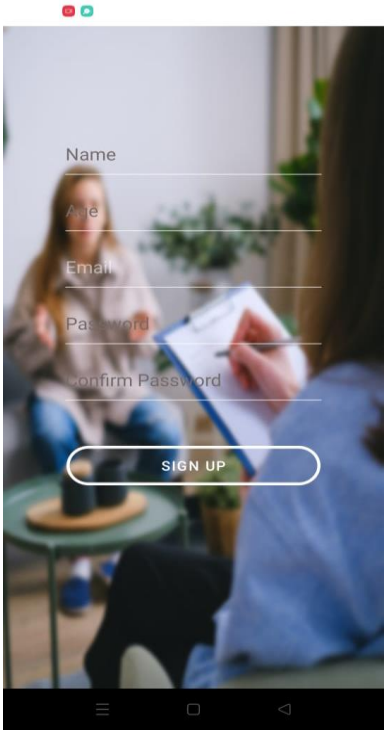


Figure 5.2: Register Page

3-Quote page:

Have only one button to enter home page.

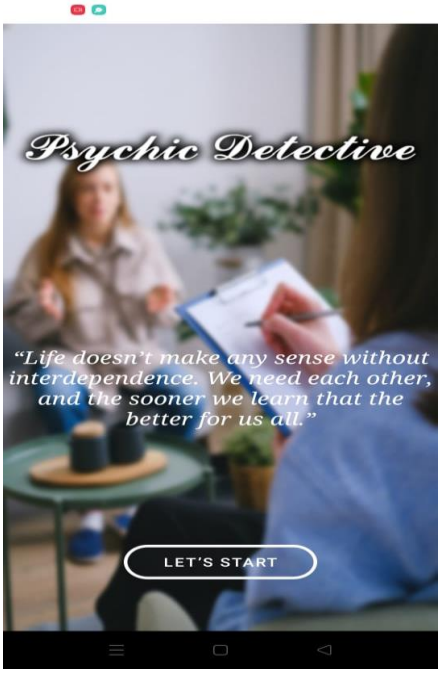


Figure 5.3 : Quote page

4-Home Page:

Contain three Images that represent type of Test:

1.Questionnaire as like in first picture in Page
2.Voice Test as like in second picture in page
3. Image Test as like in third picture in page

When press in any picture you will go to the test, Also You have Navigation Bar to allow you move between pages (go to you profile , show available doctors and logout from application) .

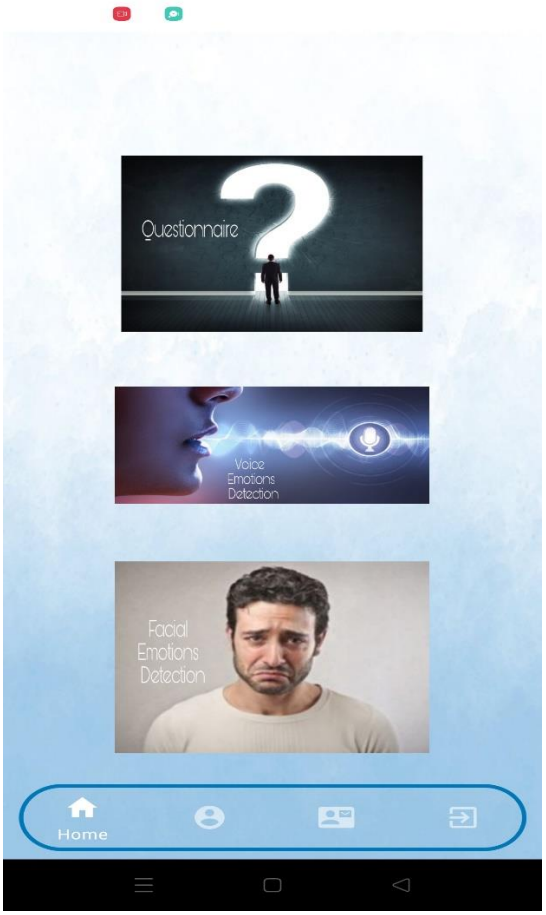


Figure 5.4: Home Page

5-Account Page:

This page show User Account information and enable to him Chang his Name , Age and Password and save this changes

Figure 5.5: Account Information Page

6- Questionnaire

There are 7 possible answers that the user can choose between, ranging from Strongly Agree (The green one) to Strongly Disagree (The red one).
Gray radio button refers to neither agree or disagree

From left to right
1- Strongly agree
2- Somewhat agree
3- A little agree
4- Neither agree or disagree
5- A little disagree
6- Somewhat agree
7- Strongly disagree

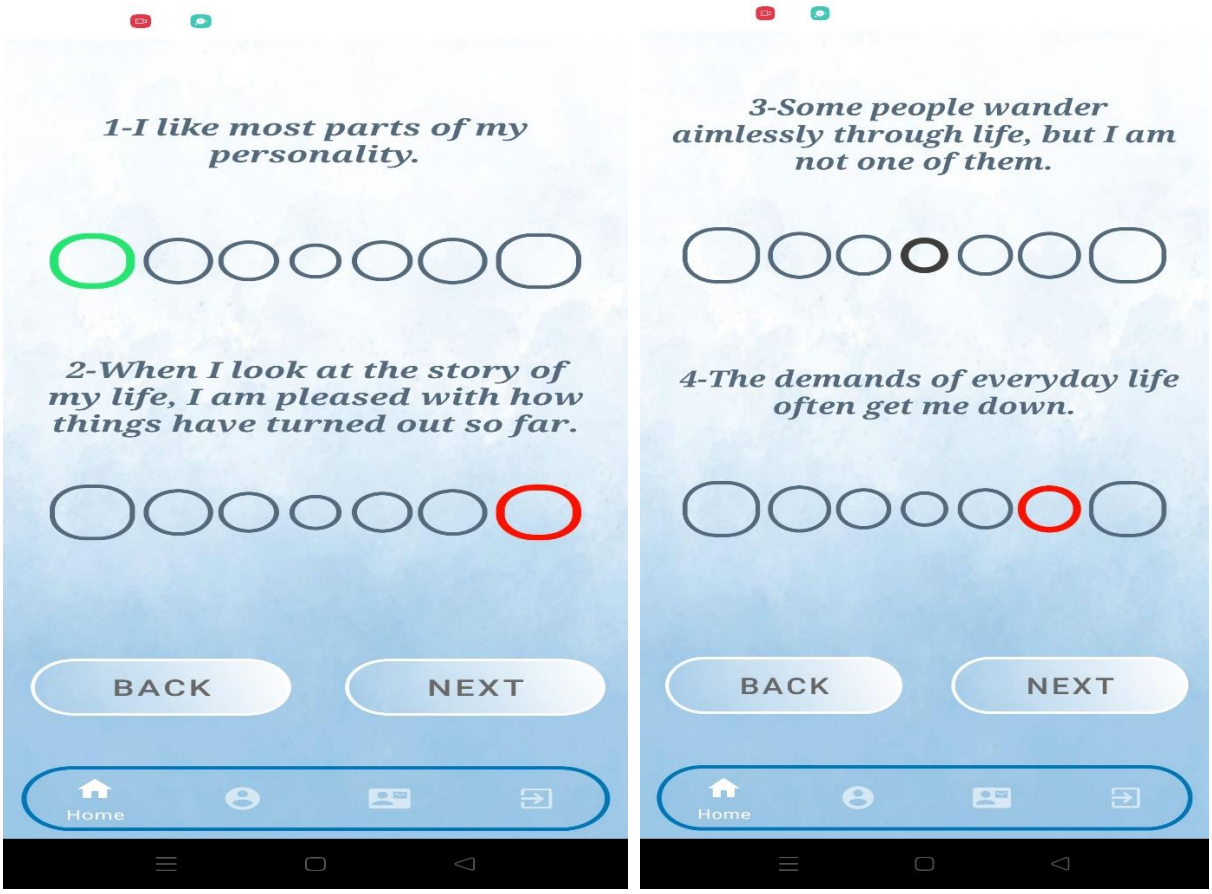


Figure 5.6: Questionnaire

7-Result of Questionnaire:

After Analyse Your answers in Questionnaire, answers represent in graph as shown in figures.

if your Score in range from 0 to 42 color of First column will change to be blue that mean your score are here as shown in figure 5.7.

if your Score in range from 43 to 83 color of Second column will change to be blue that mean your score are here as shown in figure 5.8

If your score greater than 83 color of Third column will change to be blue that mean your score are here as shown in figure 5.9

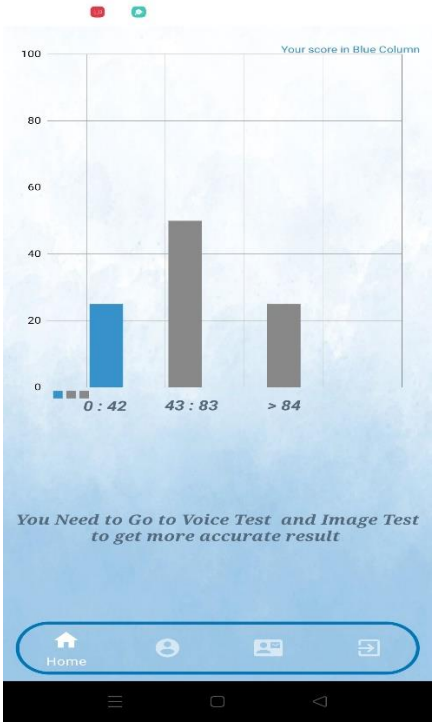


Figure 5.7: First type of Result on Questionnaire

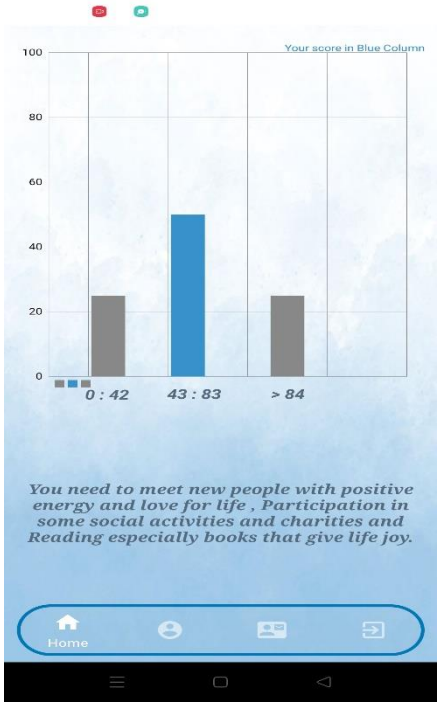


Figure 5.8: Second Type of Result on Questionnaire

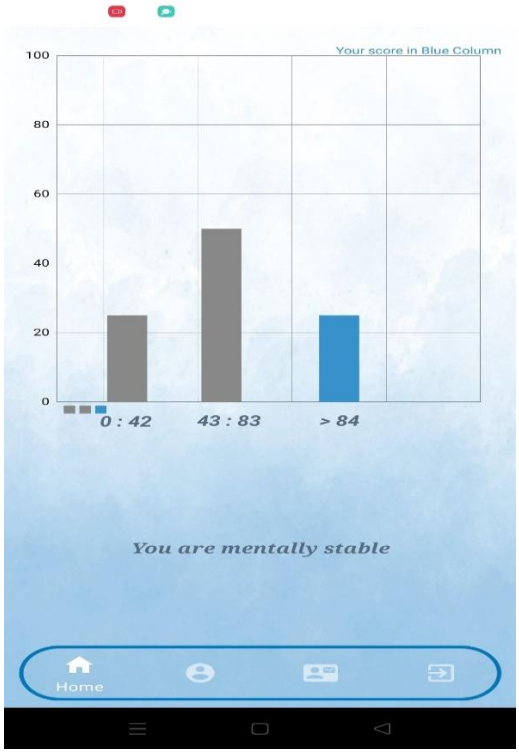


Figure 5.9: third type of Result on Questionaire

8-Voice Emotions detection:

There are 3 button, first one for start recording, second one for stop recording (ending it) and the third one for play the record and for classify the user tone in record to one of the 8 emotions

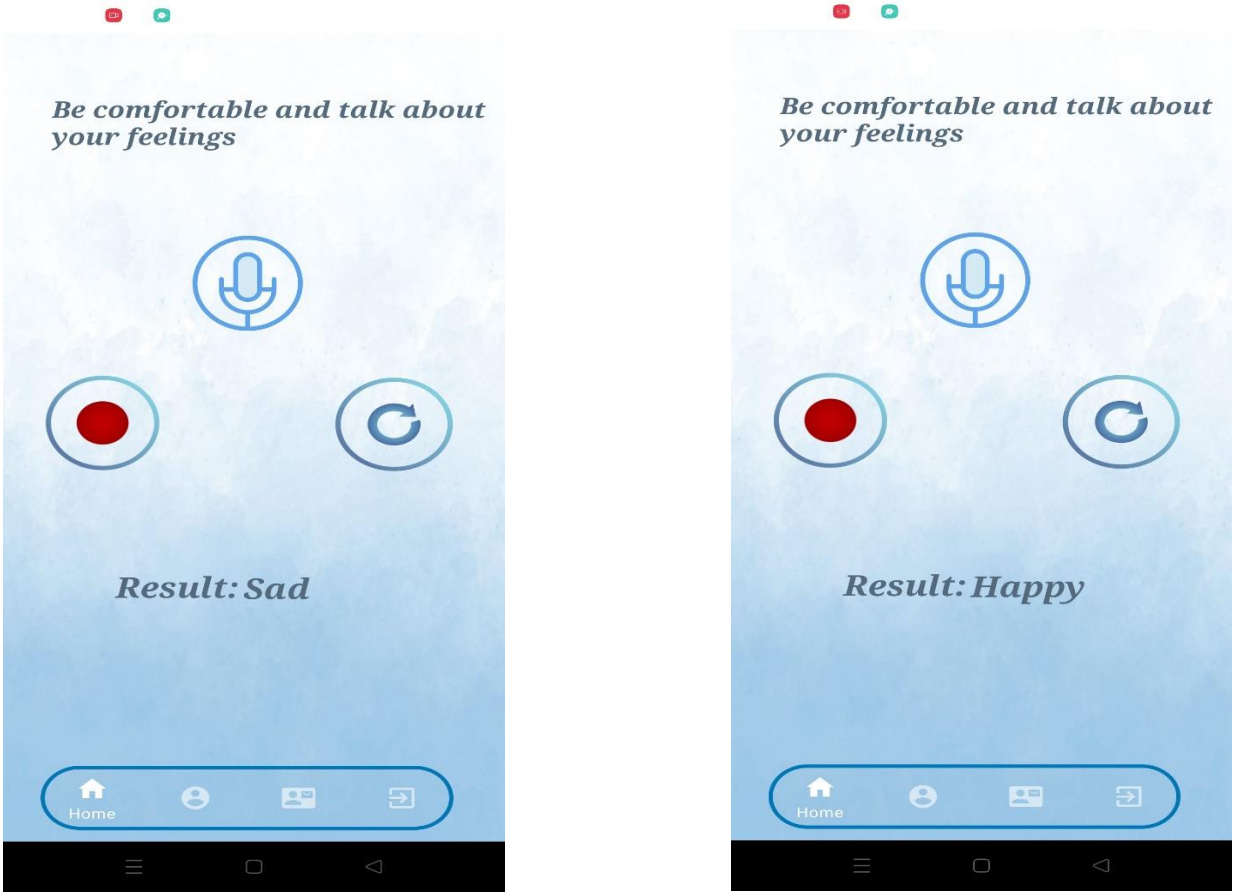


Figure 5.10: Voice result

9- Facial Expression Test:

User has to options:

- First: capture image from camera
- Second: obtain image from his gallery

Depending on his facial expression the model will classify his image to one of the 7 emotions

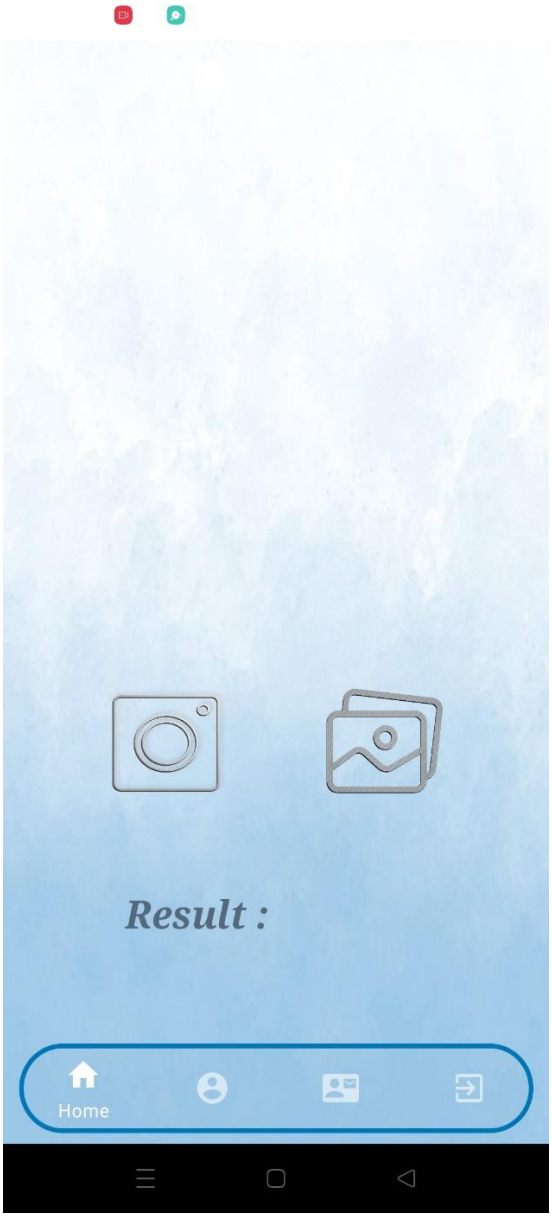


Figure 5.11: Image Test Page

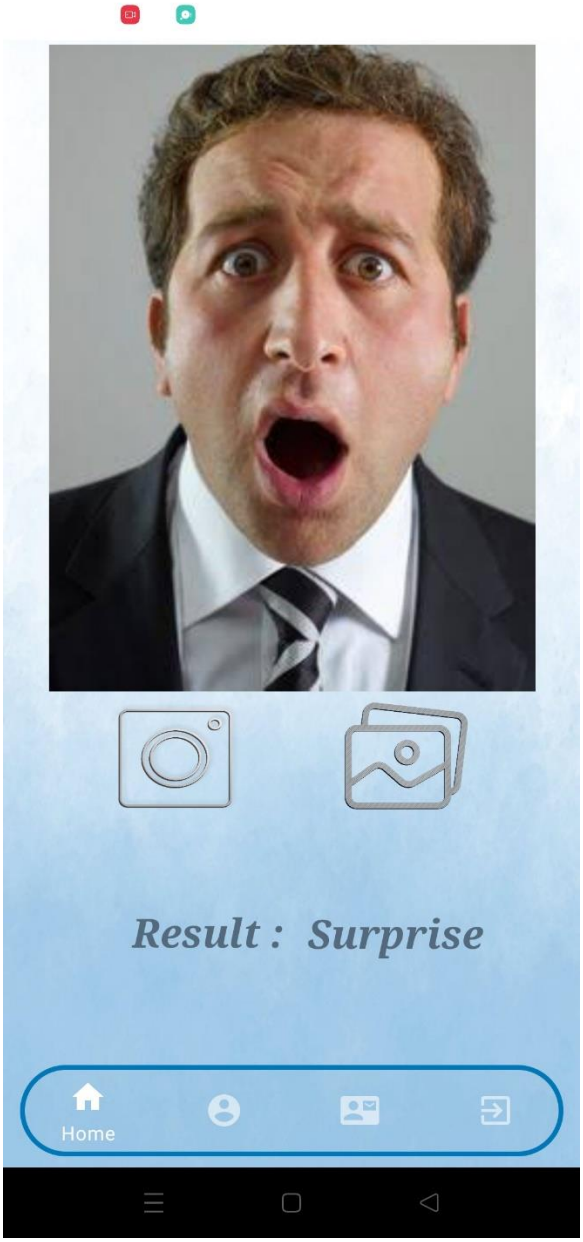


Figure 5.12: Image Test Result with surprise emotion

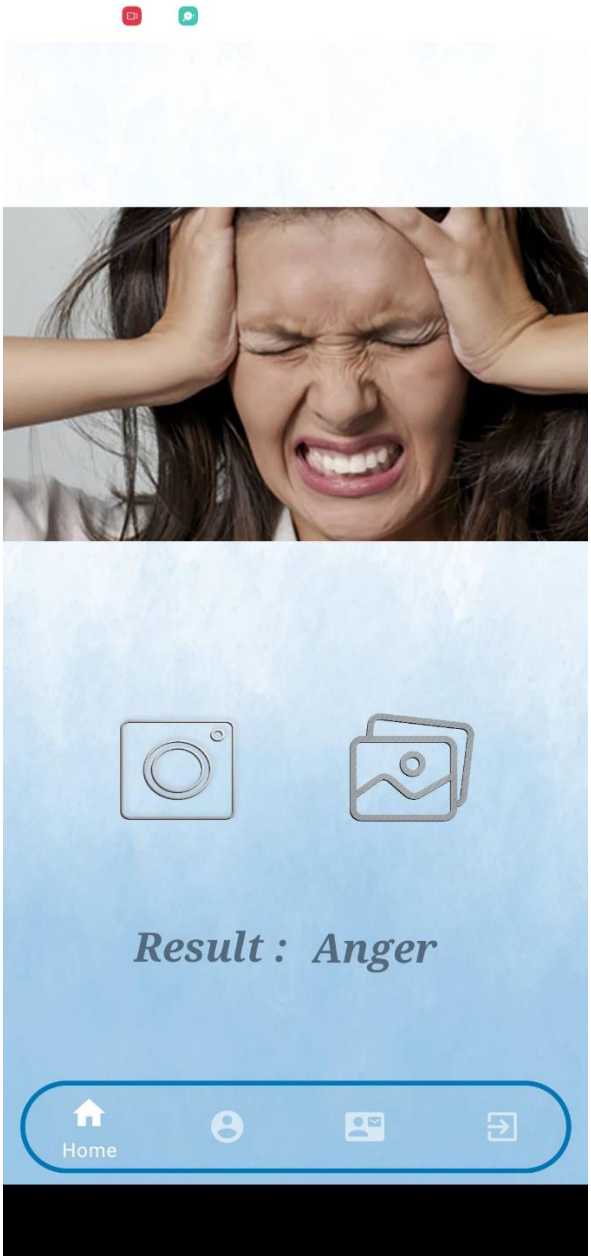


Figure 5.13Image Test Result with angry emotion

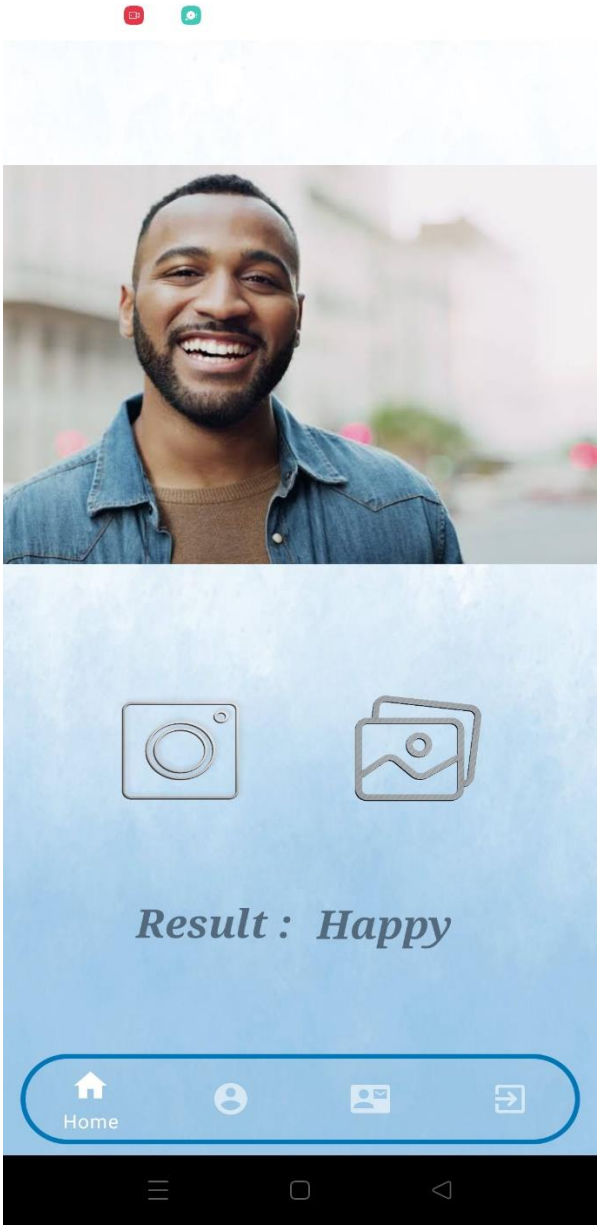


Figure 5.14 Image Test Result with happy emotion

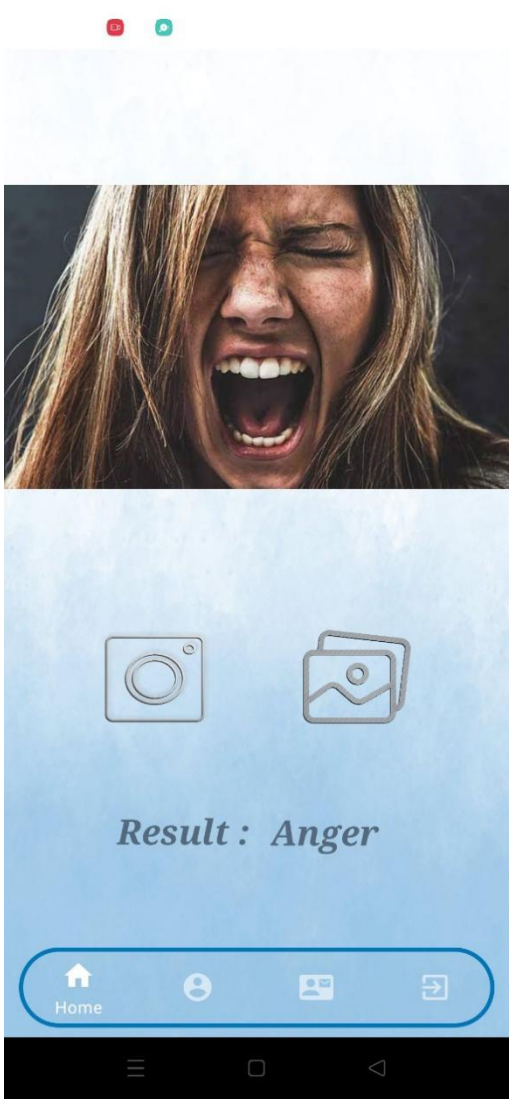


Figure 5.15 Image Test Result with angry emotion

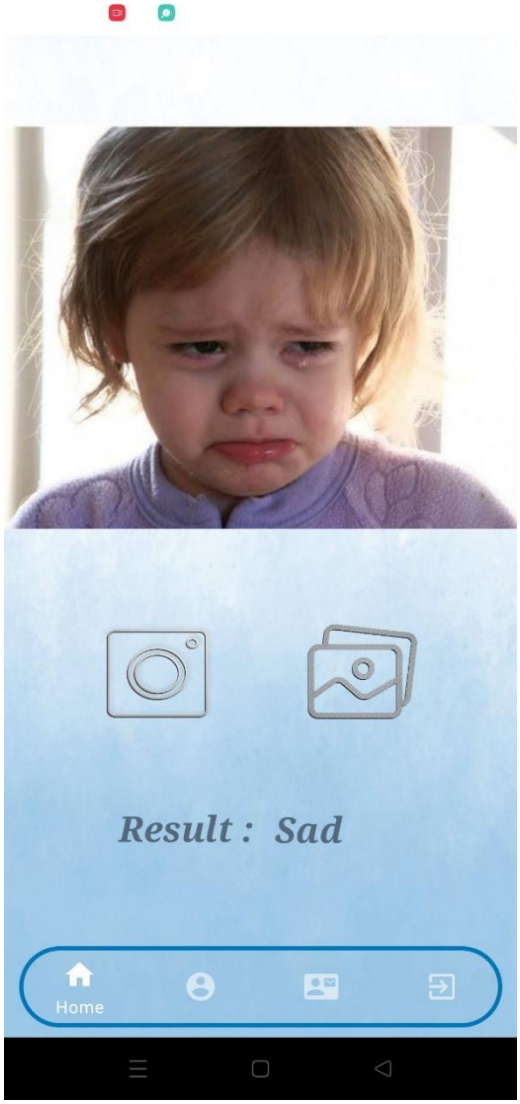


Figure 5.16 Image Test Result with sad emotion

10- Request Doctor:

It’s a list of psychiatrists with their contacts information to enable the user to contact him if he needs.

Figure 5.17: Doctor Contacts

11-Menu bar:

Contains button that allows navigation between pages

- Move to Home Page 

- Move to Account Page 
- Move to Doctors Contacts Page 
- Logout 

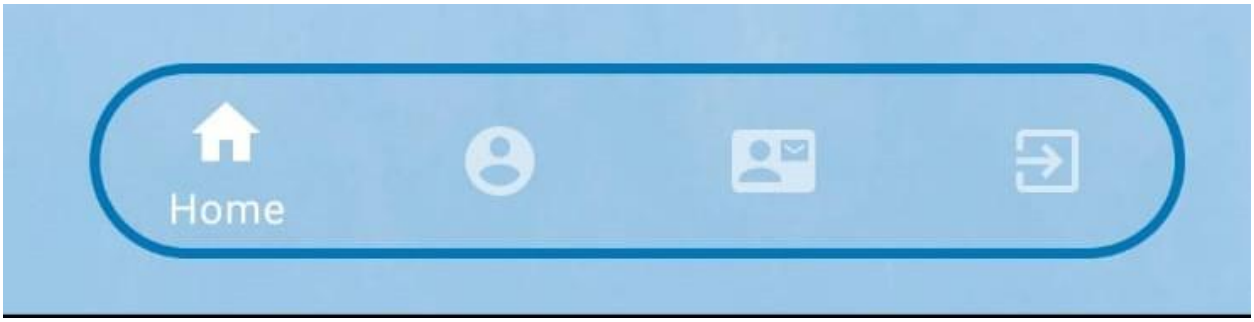


Figure 5.18: Menu Bar

Chapter 6

Conclusions and Future Work

6.1 Conclusion:

Introducing architecture for detecting emotion. The sequential model was trained on images and voices of collected datasets provided by kaggle. The images were of the dimension 48x48 pixels. The training set consists of 28,700 face images and 27,200 voice. Using same Neural Network to get feeling of person with accuracy 95% By Extracting 5 features from voice user entered

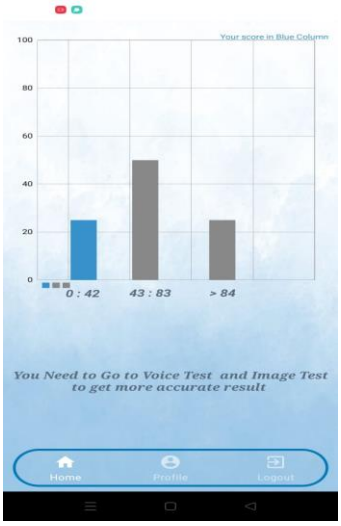


Figure 6-1

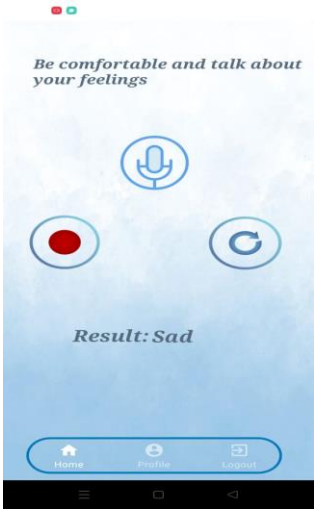


Figure 6-2

Figure 6-3

-As shown in figure this the result
Figure 6-1 result of answer question
Figure 6-2 result of voice
Figure 6-3 result after user add photo

-By extracting 5 features from voice and entered to model by using Neural Network to extract feeling from voice with accuracy 73.4%

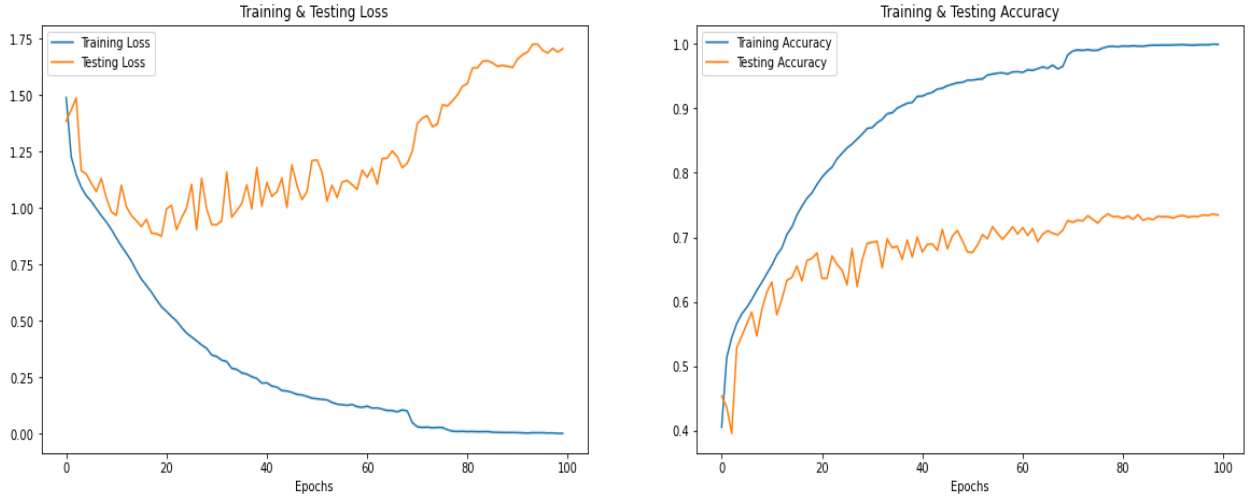


Figure 2.4: Voice Loss and Accuracy

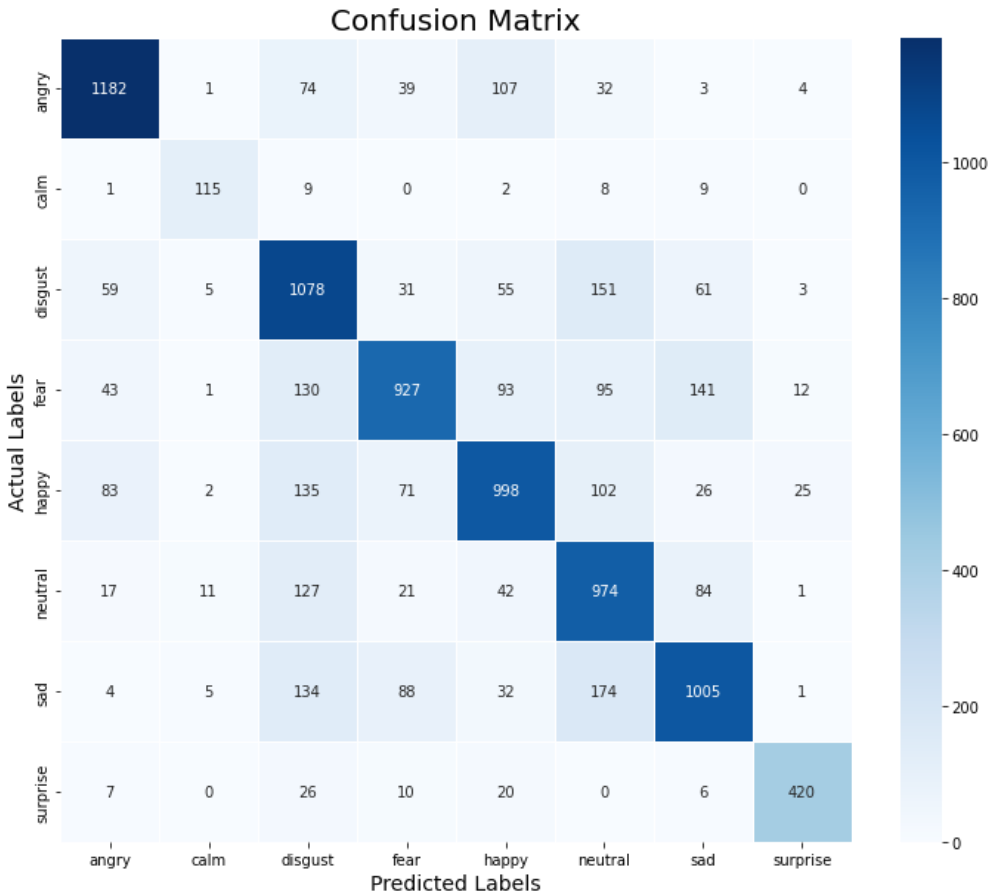


Figure 6.5: Voice Loss and Accuracy

-We use Neural Network to extract feature from image and detect feeling with accuracy 95%

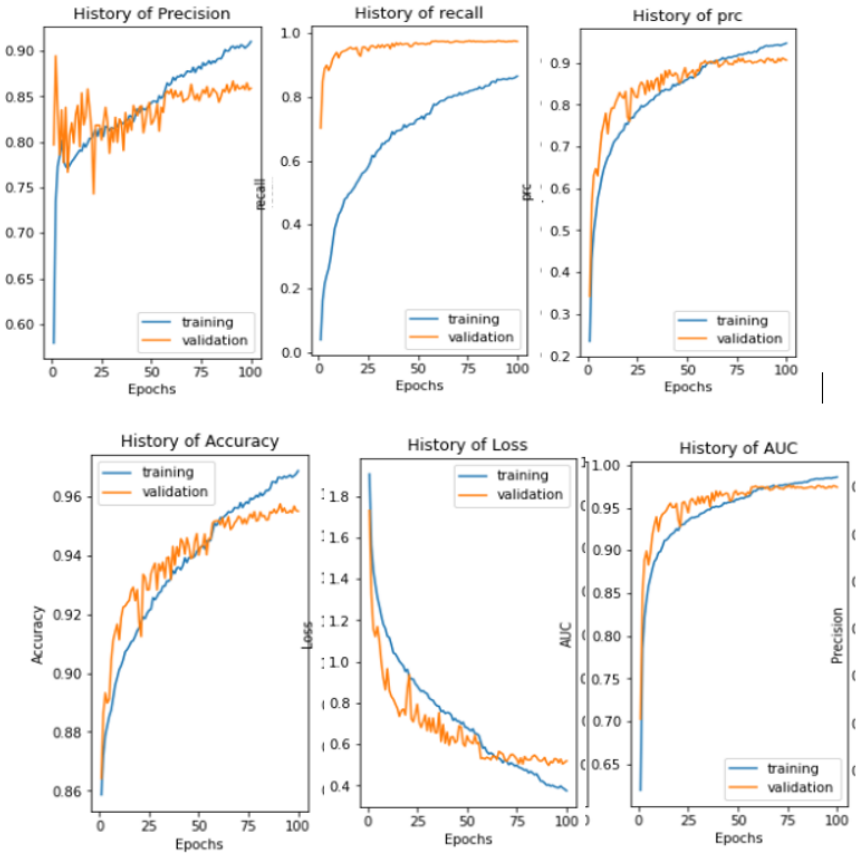


Figure 6.6: Image emotion model Evaluation

6.2 Future Work

We will provide video level and extract body language, voice from speech in video, and image snippets from it to be analyzed all at once. User can enter Voice in multiple languages and linking all levels together by the video intake with voice then speech extraction.

References

-Fernald, L. D. (2008). Psychology: Six perspectives. SAGE.

H. Alshamsi, H. Meng and M. Li, "Real time facial expression recognition app development on mobile phones," 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2016, pp. 1750-1755, doi: 10.1109/FSKD.2016.7603442.

-Higuchi M, Nakamura M, Shinohara S, et al. Effectiveness of a Voice-Based Mental Health Evaluation System for Mobile Devices: Prospective Study. JMIR Form Res. 2020;4(7):e16455. Published 2020 Jul 20. doi:10.2196/16455

Medical-Pst. MIMOSYS: Voice Analysis of Pathophysiology. 2019. Available online: https://medical-pst.com/en/ (accessed on 1 September 2019).

- Srividya, M., Mohanavalli, S. & Bhalaji, N. Behavioral Modeling for Mental Health using Machine Learning Algorithms. J Med Syst 42, 88 (2018). doi: 10.1007/s10916-018-0934-5

-D. Reney and N. Tripathi, "An Efficient Method to Face and Emotion Detection," 2015 Fifth International Conference on Communication Systems and Network Technologies, 2015, pp. 493-497, doi: 10.1109/CSNT.2015.155.

-ResearchGate: Case Studies of Utilization of the Mind Monitoring System (MIMOSYS) Using Voice and Its Future Prospects. https://www.researchgate.net/publication/320891644_Case_Studies_of_Utilization_of_the_Mind_Monitoring_System_MIMOSYS_Using_Voice_and_Its_Future_Prospects

-Becky inkster, shubhankar sarda & vinod subramanian, A Real-World Mixed Methods Data Evaluation of an Empathy-Driven, Conversational Artificial Intelligence Agent for Digital Mental Wellbeing(2018). doi: 10.2196/1210.

Alexandros Roniotis & Manolis Tsiknakis, Detecting Depression Using Voice Signal Extracted by Chatbots: A Feasibility Study (2018). **https://link.springer.com/chapter/10.1007/978-3-319-76908-0_37**

https://sci-hub.mksa.top/10.1109/fskd.2016.7603442