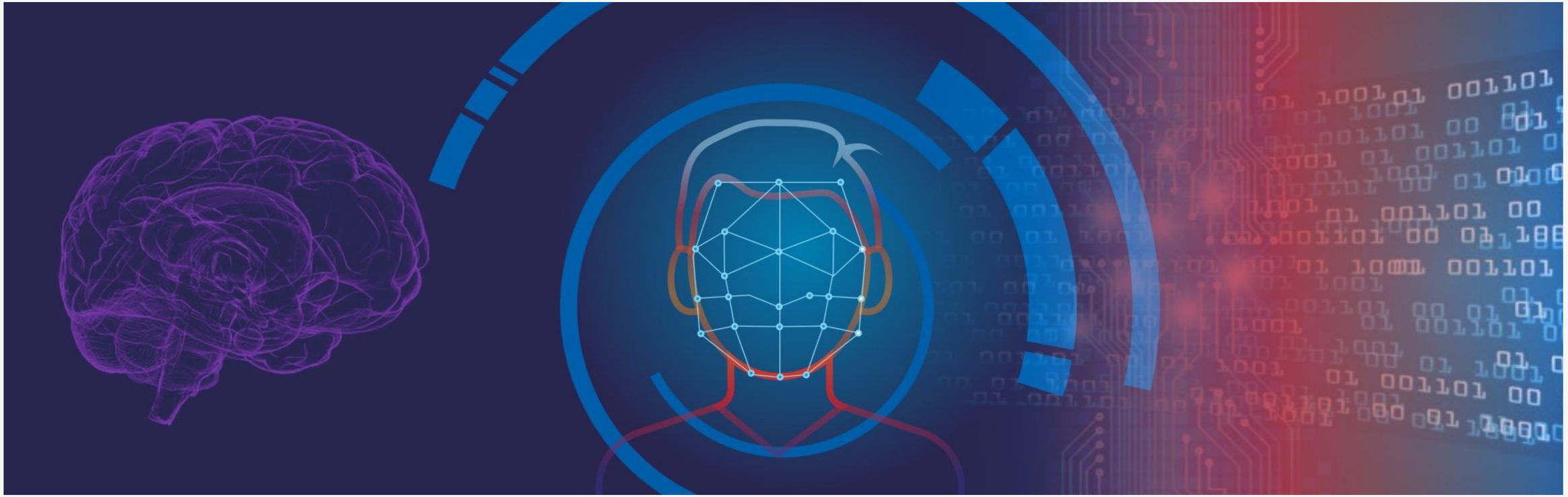


EMOTION DETECTION 2.0

MEMBERS

NAVYA MAMORIA

CHAITANYA GUPTA

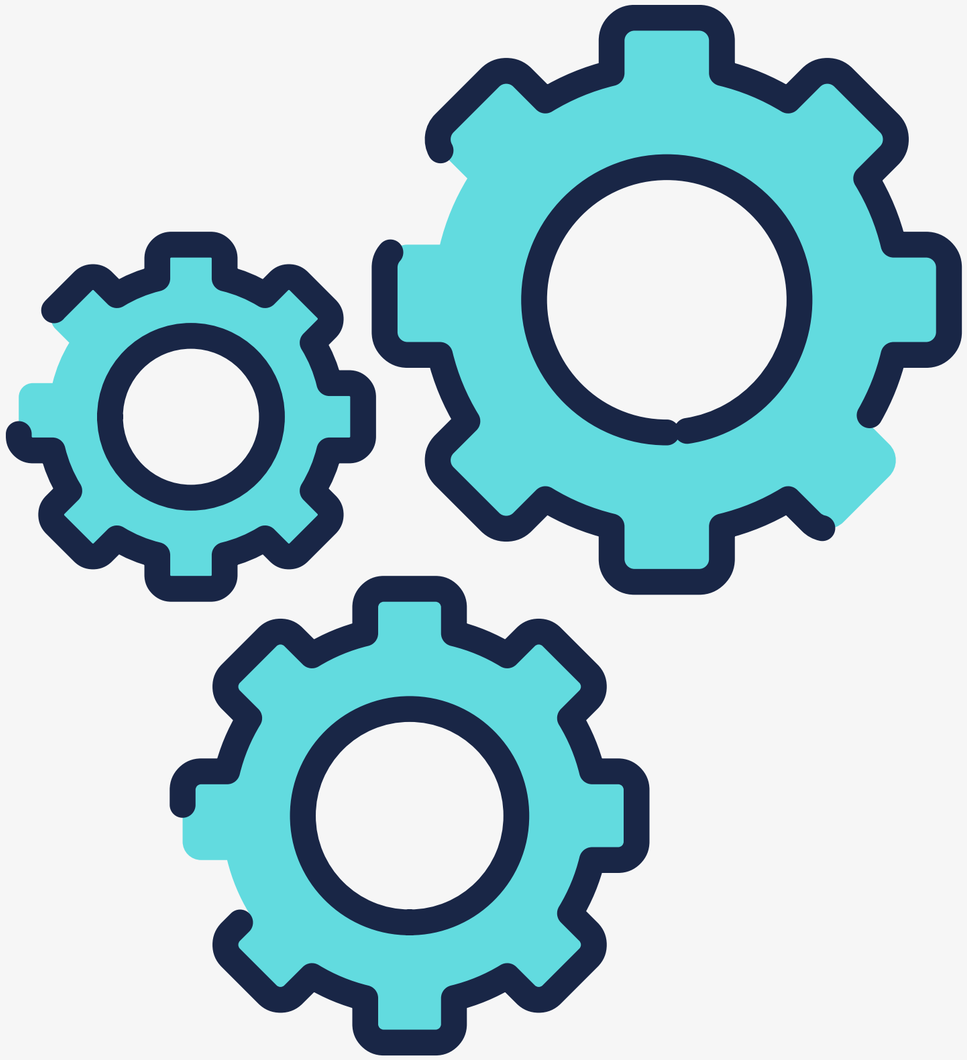
AKRITI JAIN

SURIYA R S

AAYUSH KUMAR

INTRODUCTION

- To accomplish a complete interaction between humans and machines, emotion recognition needs to be considered.
- Emotion-sensing technology can be applied in educational and diagnostic software, driverless cars, personal robots, pervasive computing, sentient virtual reality, video games, affective toys and consumer electronics devices.
- In the last five years, the field of AI has made major progress. Now it is possible to predict human emotions with much higher accuracy.



OBJECTIVE

In this project, we analyze speech and video to predict emotions.

Visual Emotion Detection

DATASET USED

- **FER 2013**

The data consists of 48x48 pixel grayscale images of faces. The faces have been automatically registered so that the face is more or less centered 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.

REQUIREMENTS

- Open CV
- Numpy
- Keras
- Matplotlib
- Google Colab

MODEL USED

- Inception ResNet v2
- Using TensorFlow backend.

OUR WORKFLOW

Loading the dataset

Mount the datasets on Google drive and unzip them separately.

Testing and Partition of image files

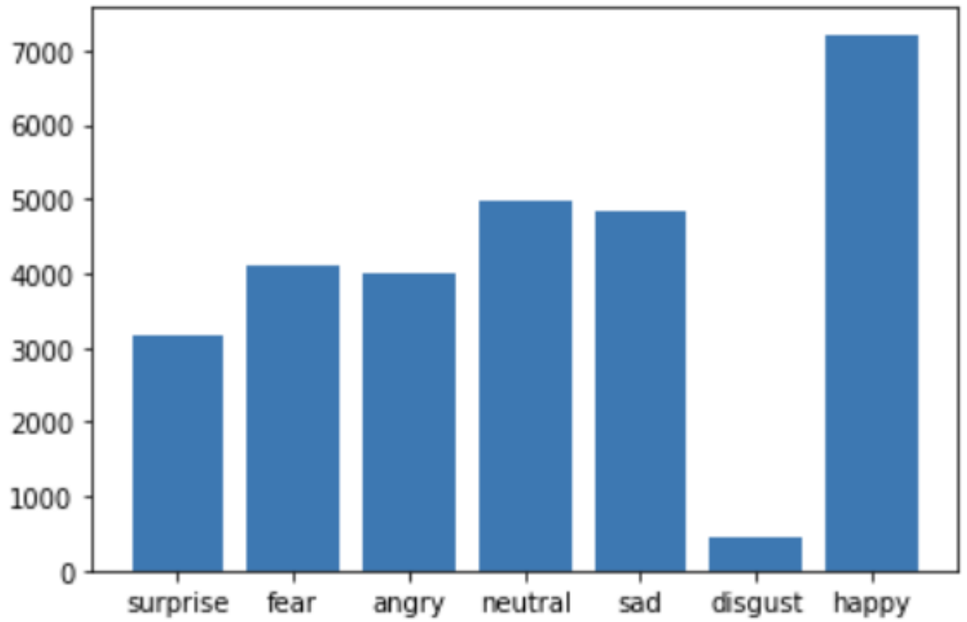
we are loading data and splitting them into train validation and test data set.

ANGRY

[4]:
```
plot_images(train_path+'/angry')
```

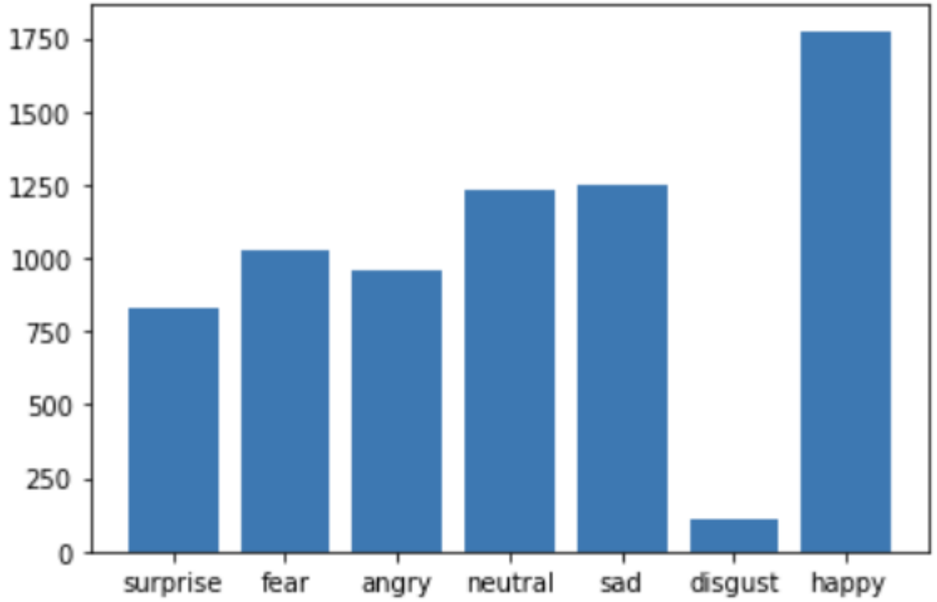
DISGUSTED

5]:
```
plot_images(train_path+'/disgust')
```



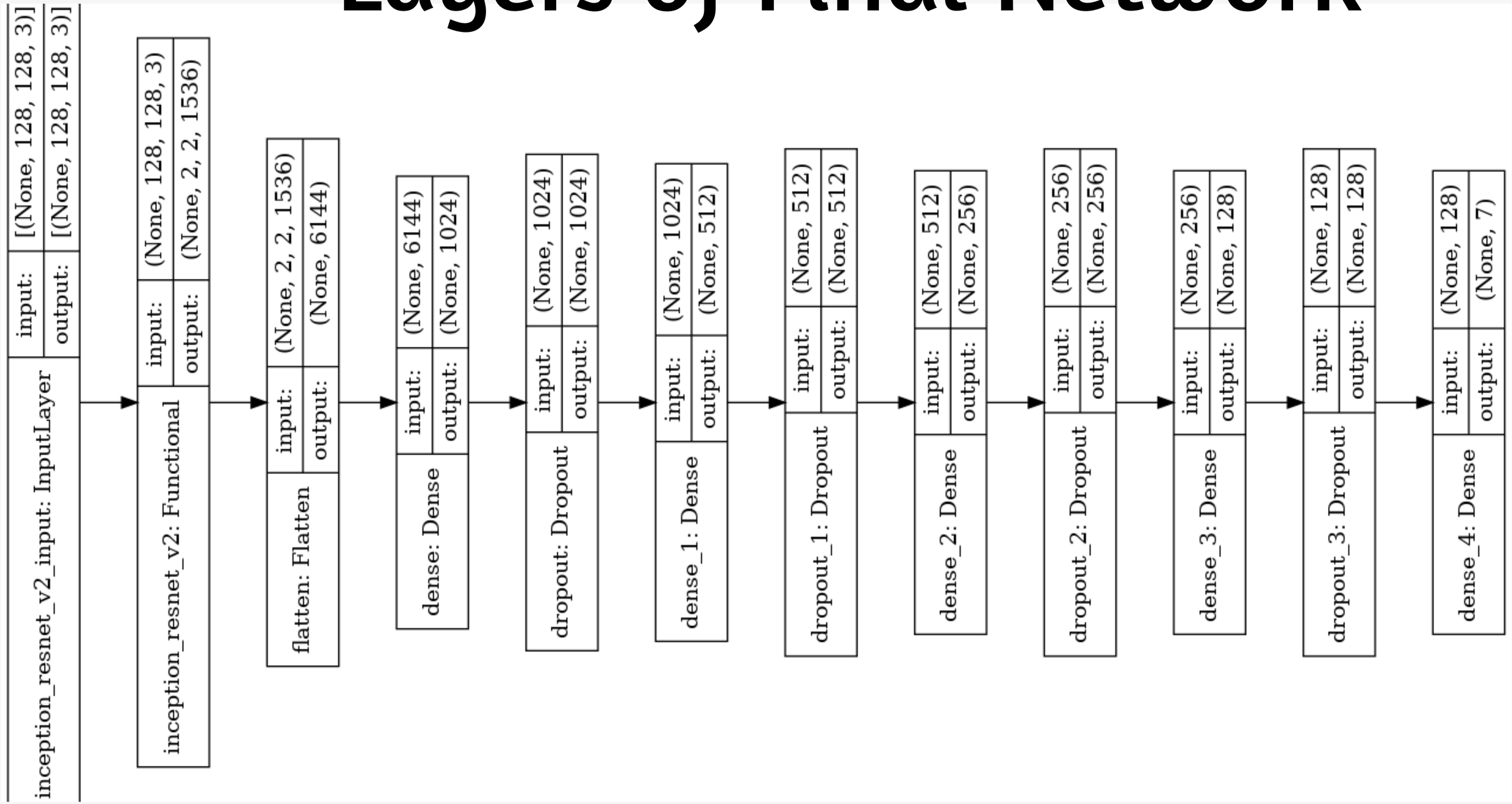
Training set

A total of 35887 images are present in the dataset. Around 22000 images is being used for training and for validation we are using around 6000 images and for testing it is around 7000.

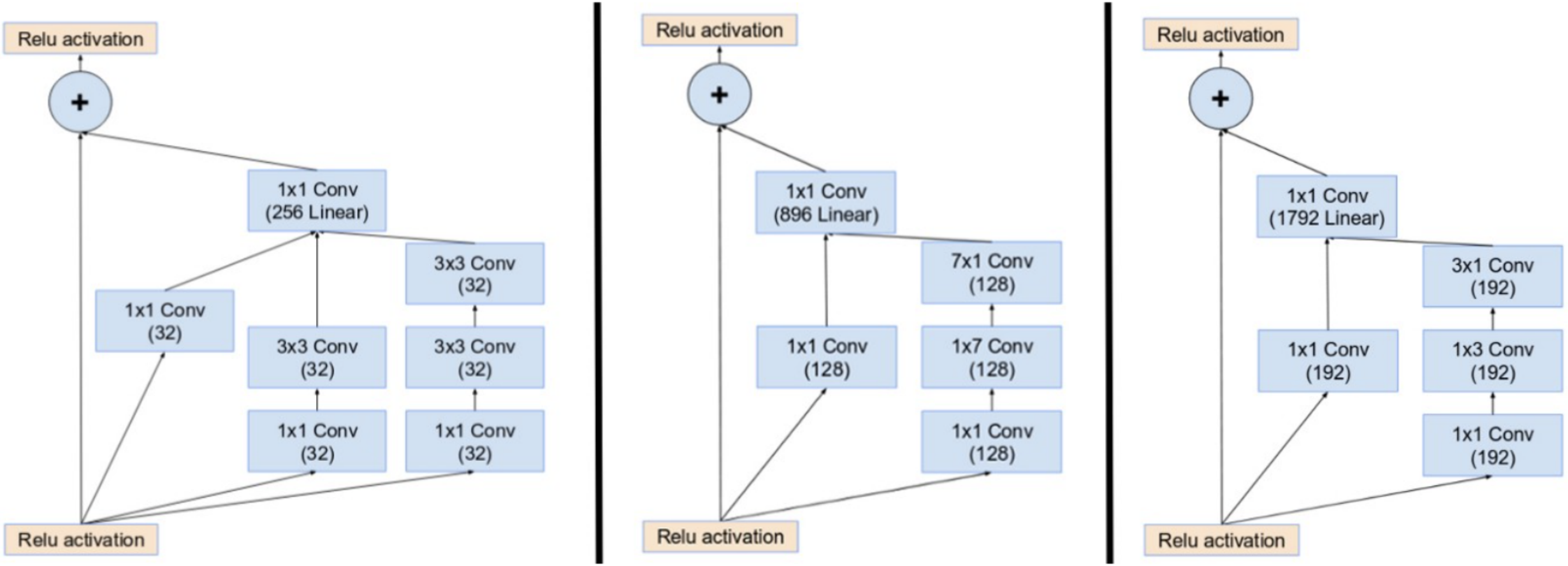


Validation set

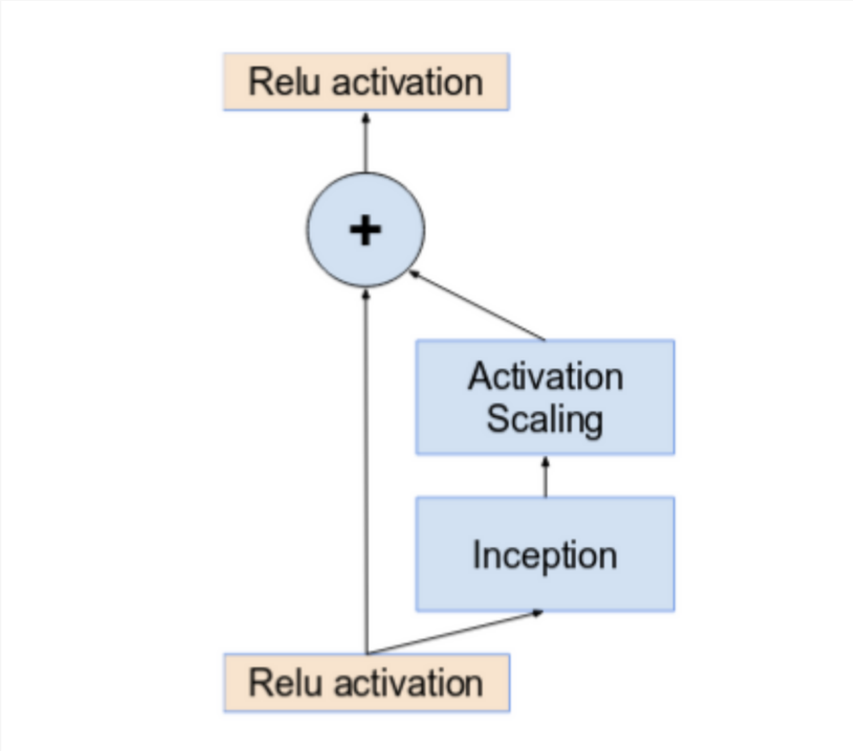
Layers of Final Network



InceptionResNet Module

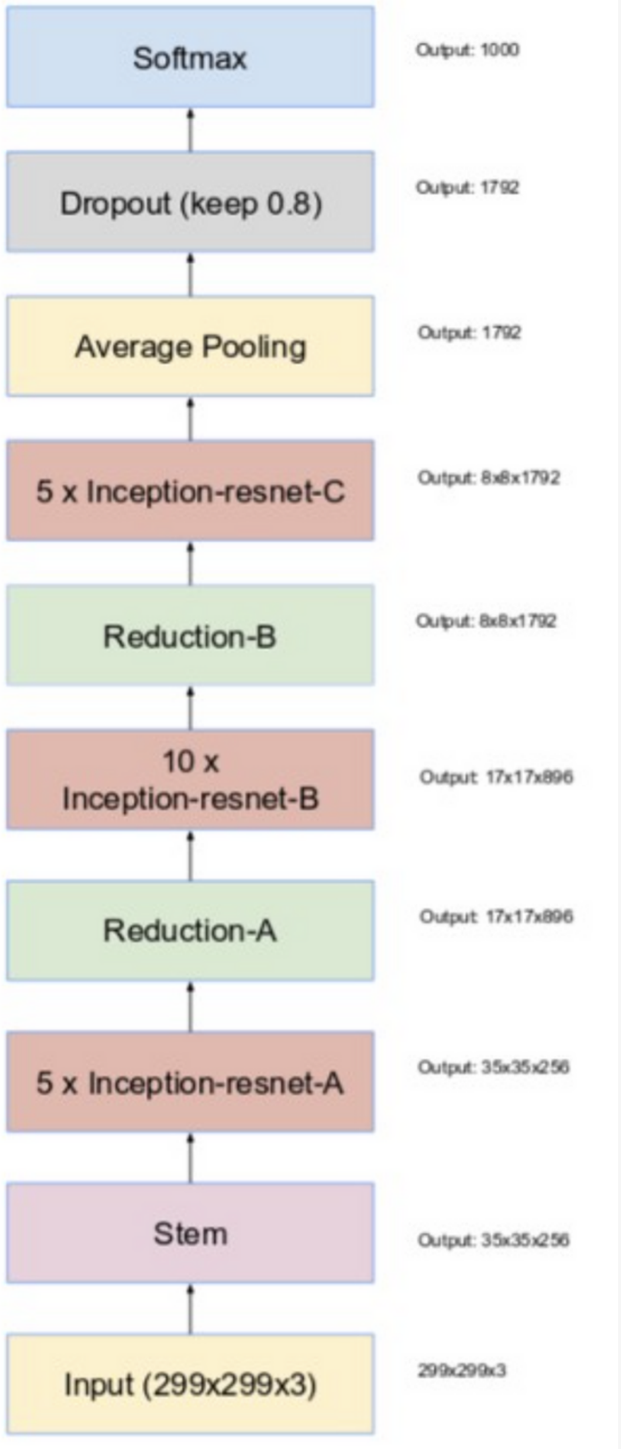


(From left) Inception modules A,B,C in an Inception ResNet. Note how the pooling layer was replaced by the residual connection, and also the additional 1×1 convolution before addition. (Source: Inception v4)



Reduction Blocks”

which are used to change the width and height of the grid



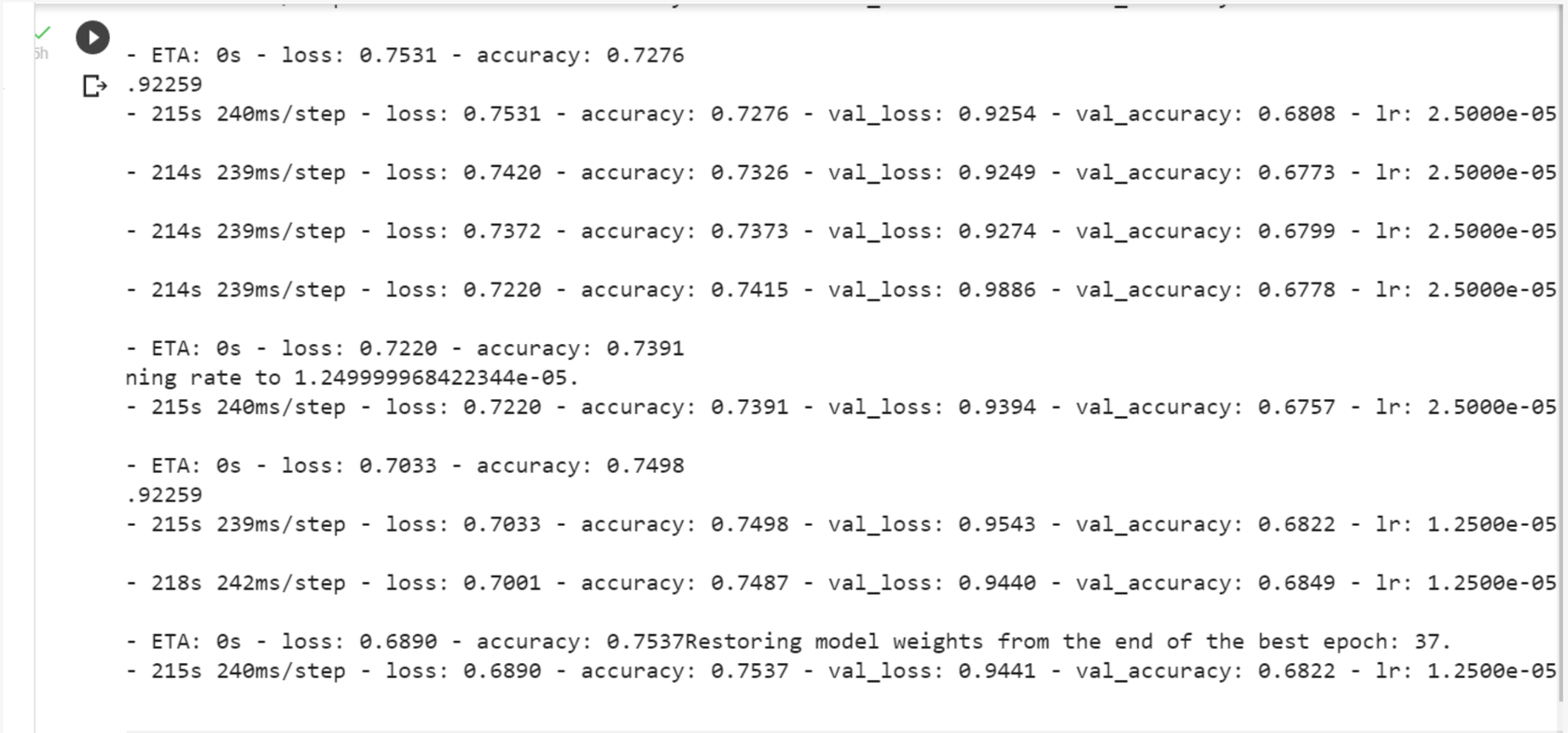
Available models

Model	Size (MB)	Top-1 Accuracy	Top-5 Accuracy	Parameters	Depth	Time (ms) per inference step (CPU)	Time (ms) per inference step (GPU)
Xception	88	79.0%	94.5%	22.9M	81	109.4	8.1
VGG16	528	71.3%	90.1%	138.4M	16	69.5	4.2
VGG19	549	71.3%	90.0%	143.7M	19	84.8	4.4
ResNet50	98	74.9%	92.1%	25.6M	107	58.2	4.6
ResNet50V2	98	76.0%	93.0%	25.6M	103	45.6	4.4
ResNet101	171	76.4%	92.8%	44.7M	209	89.6	5.2
ResNet101V2	171	77.2%	93.8%	44.7M	205	72.7	5.4
ResNet152	232	76.6%	93.1%	60.4M	311	127.4	6.5
ResNet152V2	232	78.0%	94.2%	60.4M	307	107.5	6.6
InceptionV3	92	77.9%	93.7%	23.9M	189	42.2	6.9
InceptionResNetV2	215	80.3%	95.3%	55.9M	449	130.2	10.0
MobileNet	16	70.4%	89.5%	4.3M	55	22.6	3.4
MobileNetV2	14	71.3%	90.1%	3.5M	105	25.9	3.8
DenseNet121	33	75.0%	92.3%	8.1M	242	77.1	5.4
DenseNet169	57	76.2%	93.2%	14.3M	338	96.4	6.3
DenseNet201	80	77.3%	93.6%	20.2M	402	127.2	6.7

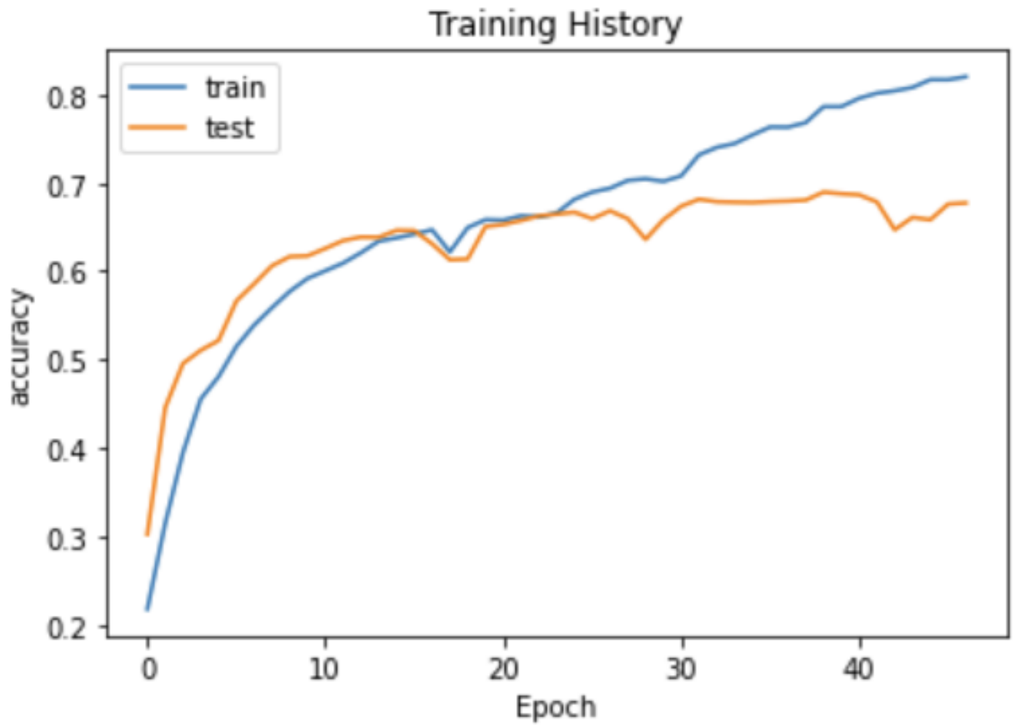
Training the Model

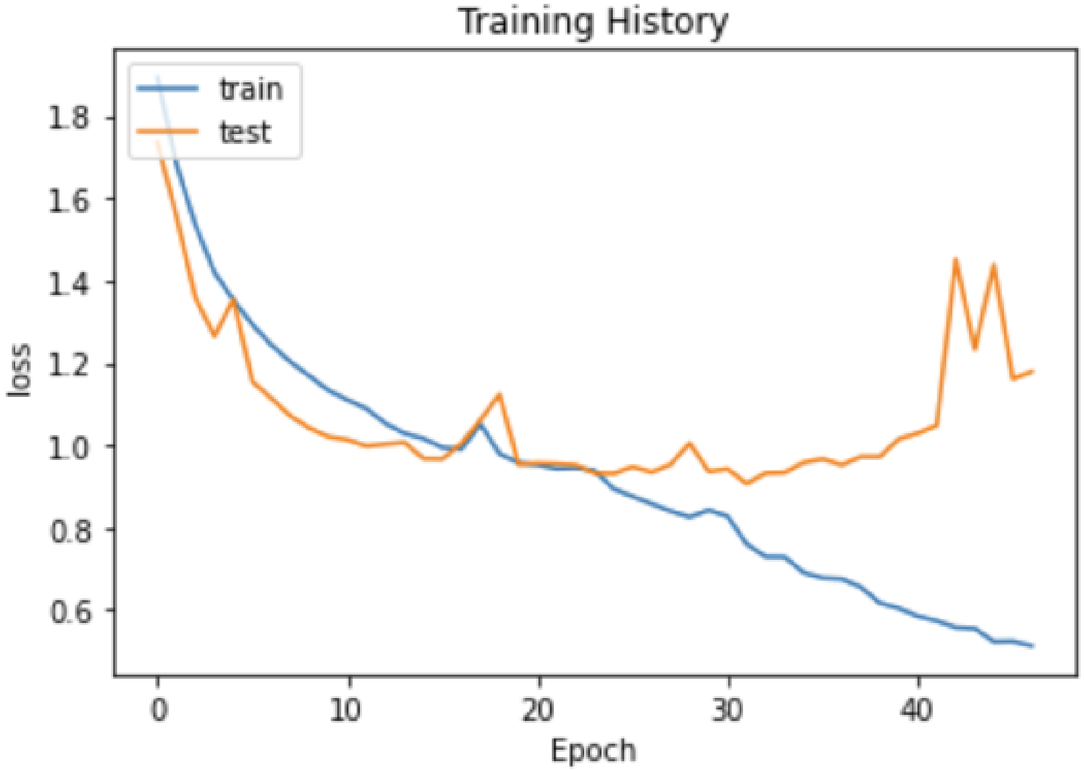
For training we are sending data in a batches of 32 /64 for making our back propogation run faster . We are using ADAM as our optimizer instead of RMSprop or momentum. After completing a total of 52 epochs we arrived at around 69% validation accuracy and 75% train accuracy.

The benchmark accuracy for FER2103 dataset is 76.82% till date and it is done by ensemble ResmaskingNet with 6 other CNNs .

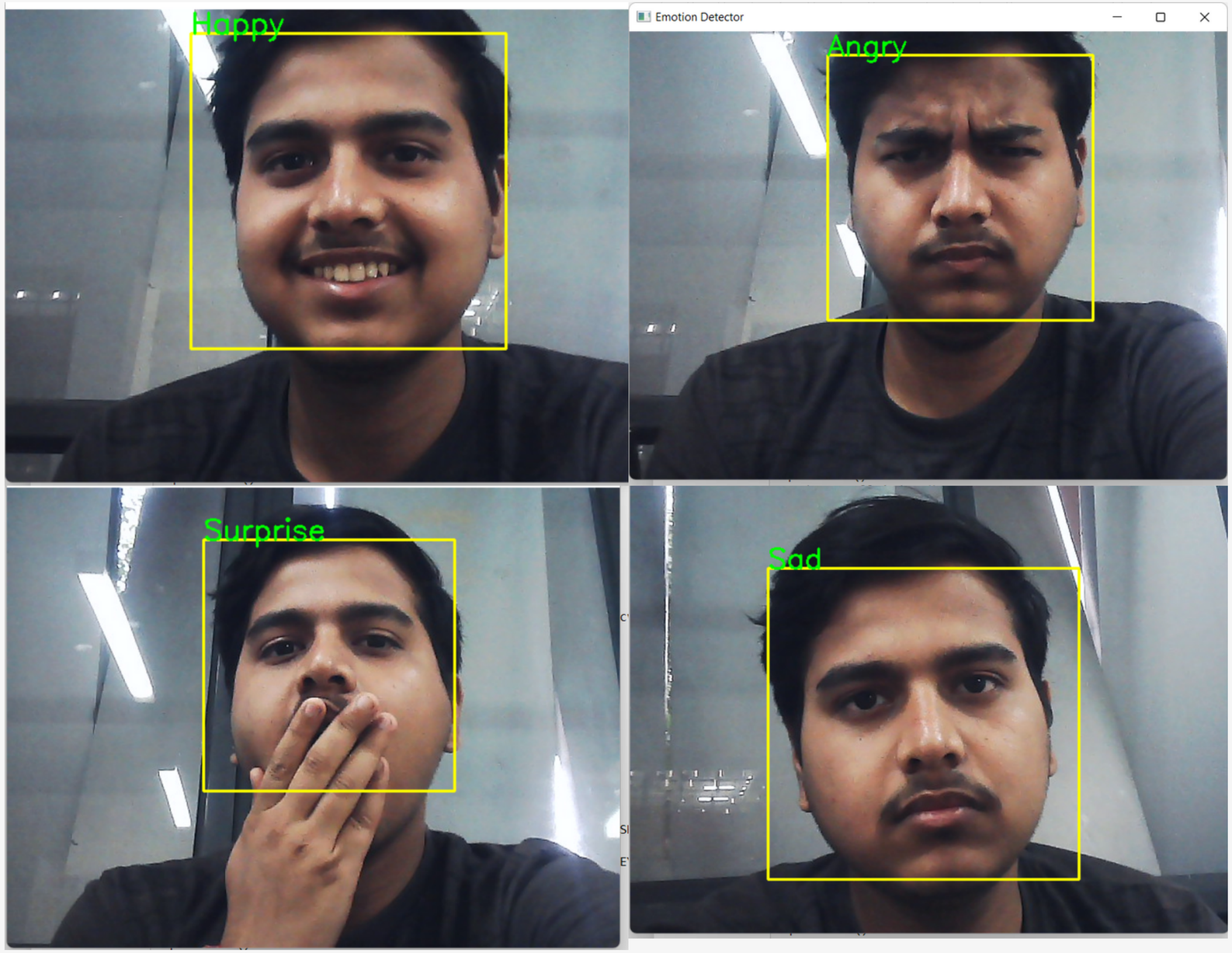


Training and Validation Results .





Real Time Emotion Detection Results



FUTURE PROSPECTS

- Using different datasets
- concatenation of the result we get from audio and visuals part using multimodal
- using different techniques such as data augmentation, oversampling,undersampling etc

Audio Emotion Detection

DATASET USED

- **RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)**

The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust emotions, and song contains calm, happy, sad, angry, and fearful emotions.

- Speech file contains 1440 files: 60 trials per actor x 24 actors = 1440.
- Song file contains 1012 files: 44 trials per actor x 23 actors = 1012.

- **TESS**

TESS contains 2800 files. A set of 200 target words were spoken in the carrier phrase "Say the word _____' by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral).

REQUIREMENTS

- Sci-kit learn
- Librosa
- Keras
- Google Colab

MODEL USED

- 1D CNN and/or LSTM
  Using TensorFlow backend.

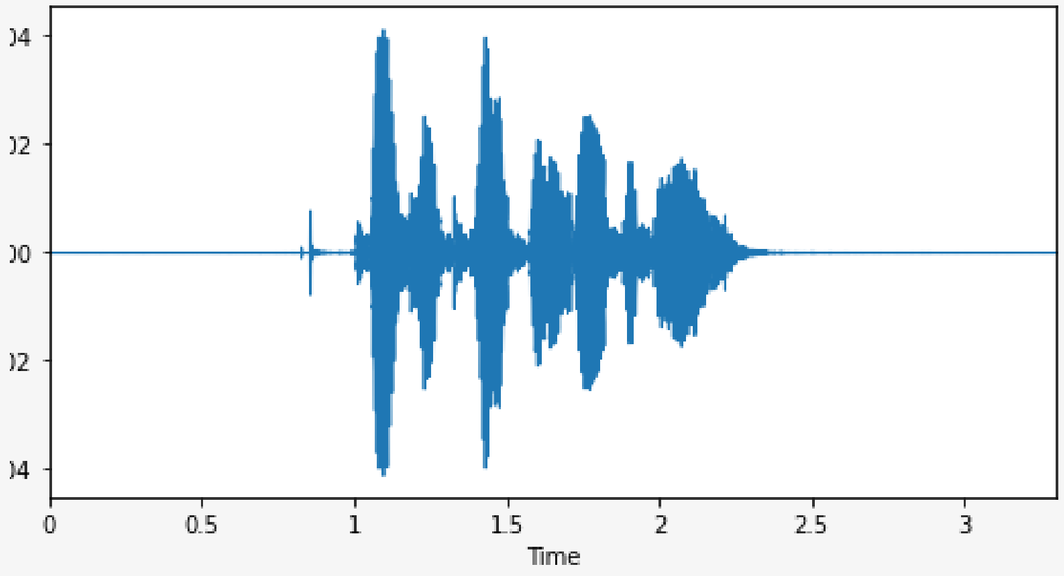
OUR WORKFLOW

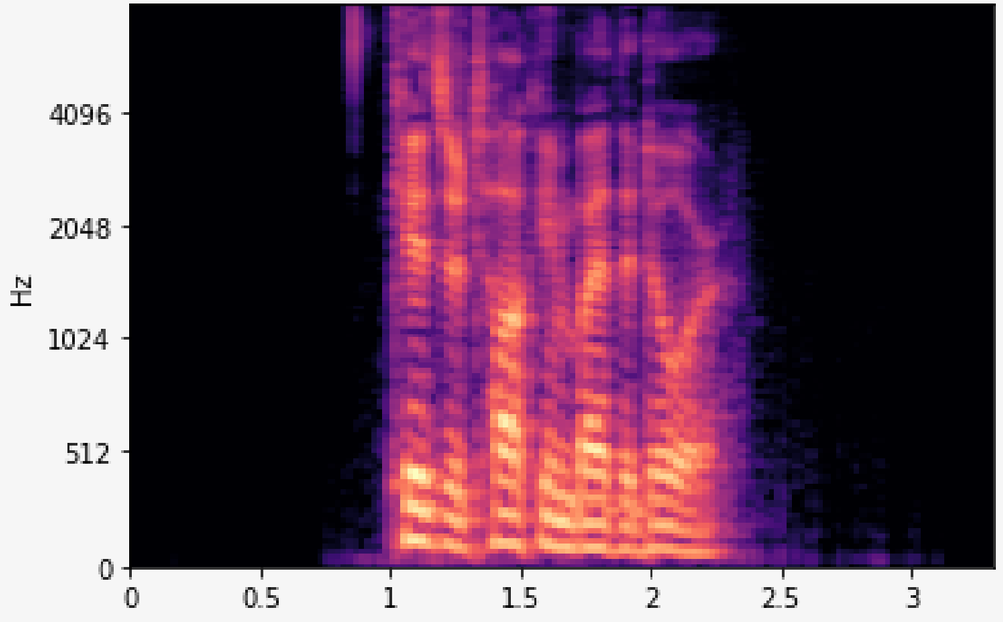
Step 1- Loading the dataset.

Firstly, we mounted the datasets on Google drive and unzipped them separately.

Step 2- Testing audio files

Tested one audio file by plotting its waveform and spectrogram.





Step 3- Feature Extraction

The next step involves extracting the features from the audio files, which will help in identifying the components of the audio signal that can easily help to distinguish emotions embedded. For feature extraction, we use the LibROSA library in python, one of the libraries used for audio analysis.

Features extracted- MFCC(Mel-Frequency Cepstral Coefficients) AND Mel spectrogram.

Step 4- Creating a dataframe

Creating a final data frame consisting of features and emotions.

Step-5 Splitting the data into test and train sets.

Step-6 Data Preprocessing

Using MinMaxScaler and label encoding emotions of y train and y test set.

Step-7 Building 1D CNN model

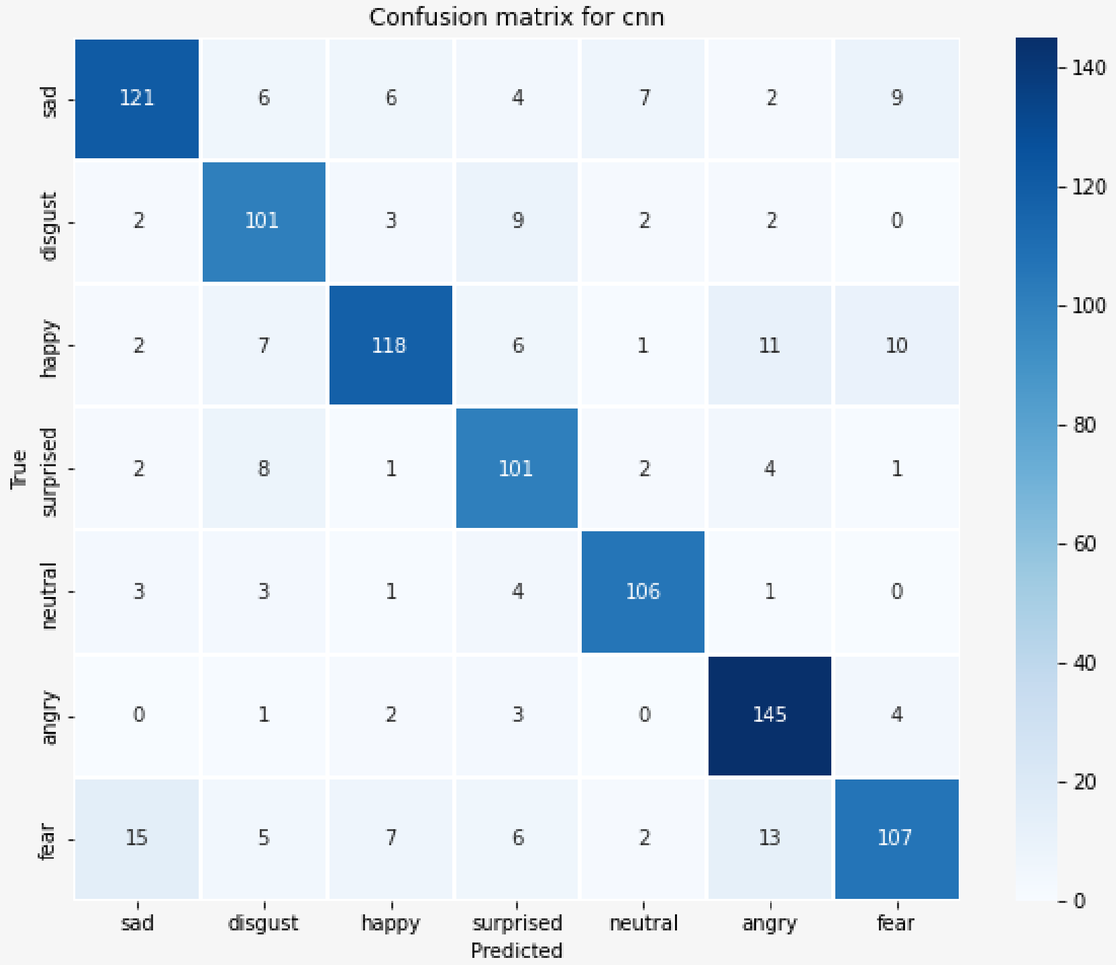
We built A 1D CNN model with 12 layers-used ReLu function as our activation function for 4 NN layers and a final layer of NN as a softmax classifier. We also added three dropout layers to prevent overfitting.

Step-8 Prediction

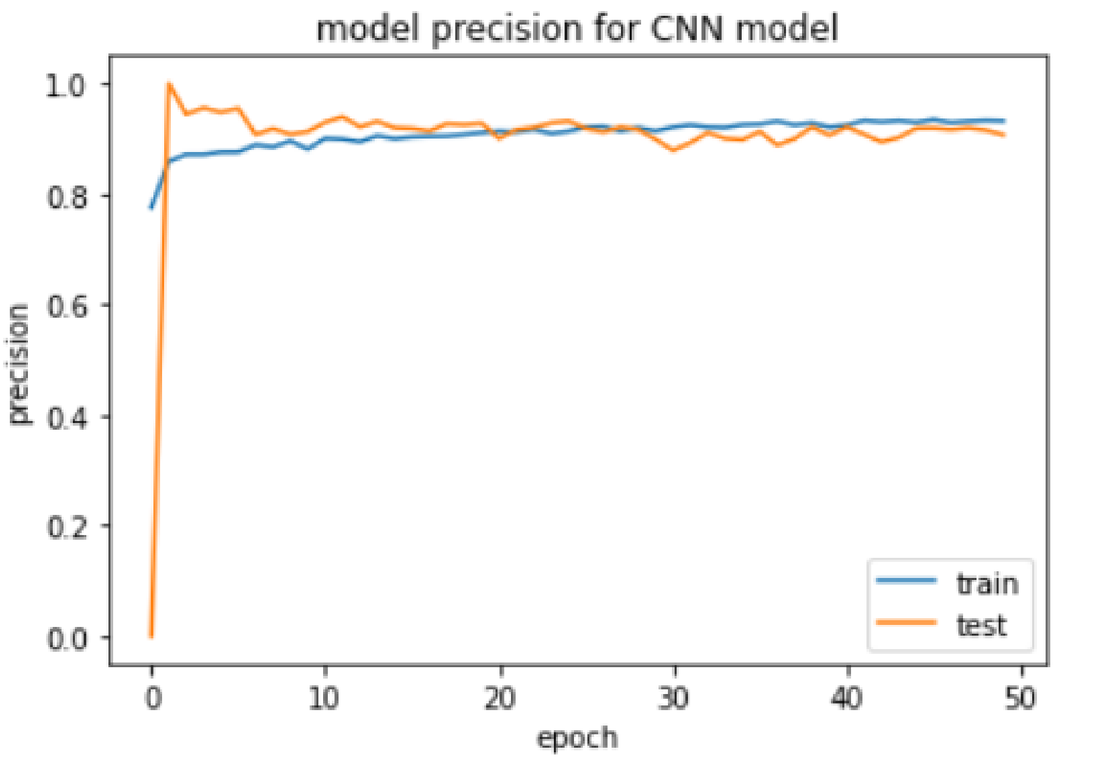
Predicting the emotions of the Y test.

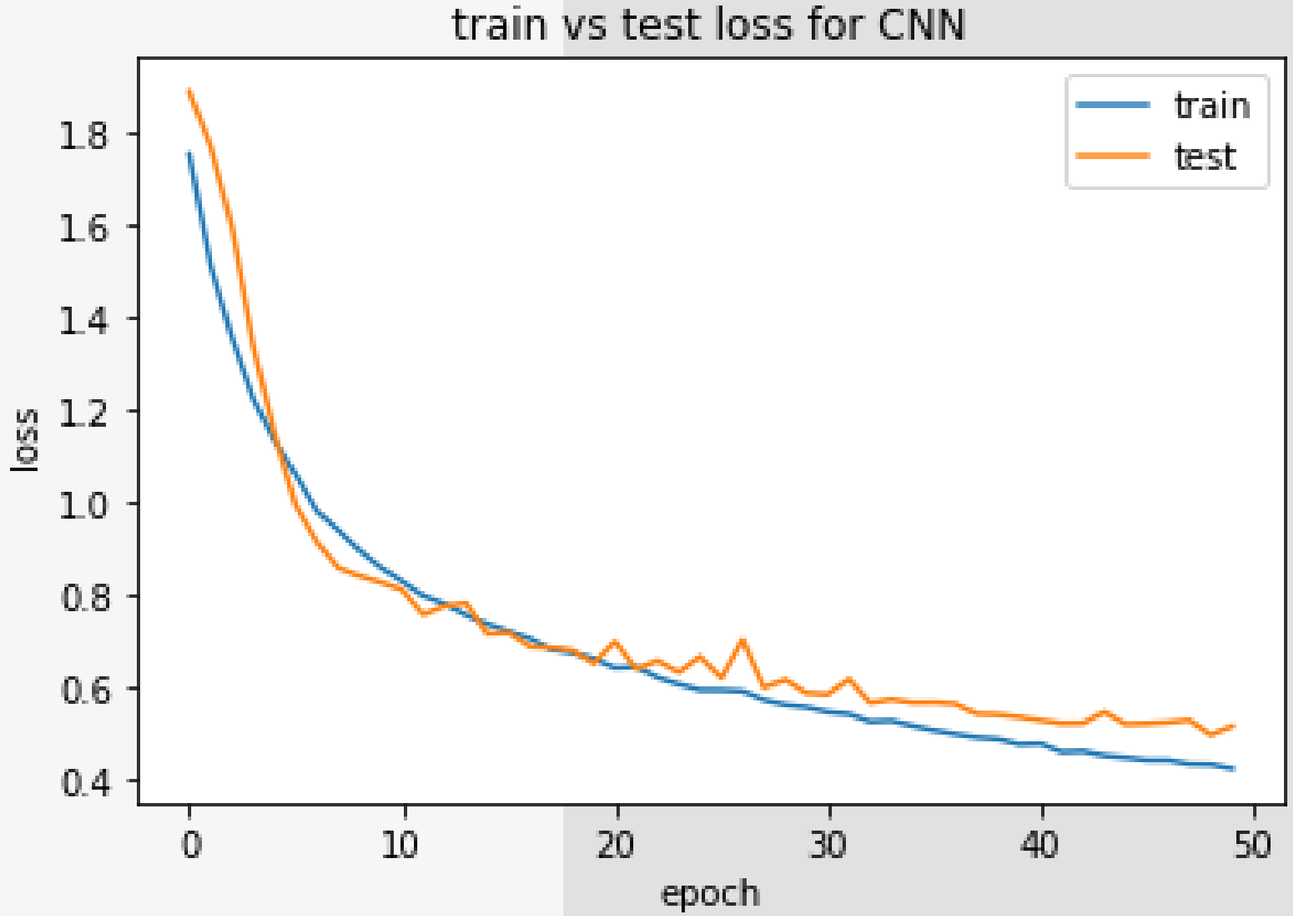
Step-9 Confusion Matrix

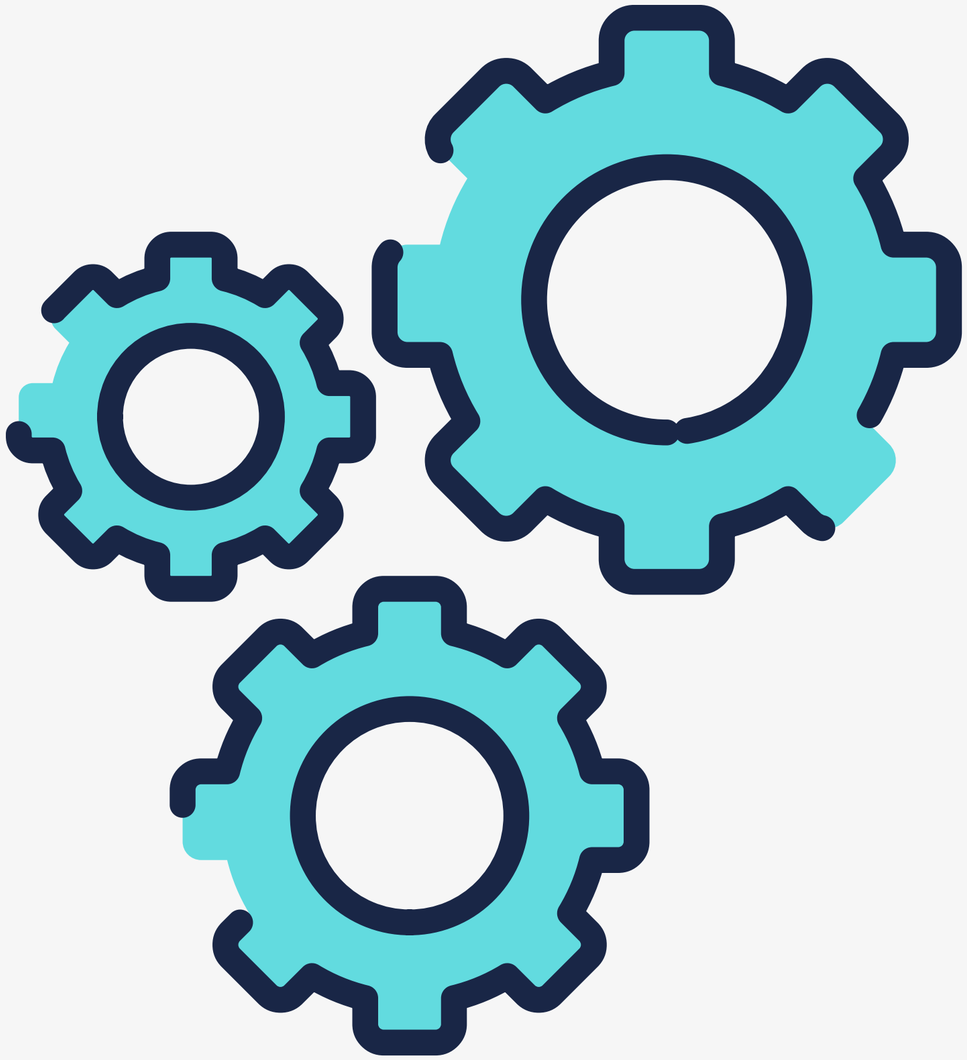
For checking the model's accuracy and finding the number of true positives, true negatives, false positives, and false negatives.



We used Adam as an optimizer .







CONCLUSION

- The 1D CNN model gave an F1 score of 81% for testing seven emotions that are **happy, sad, neutral, disgust, angry, fear, and surprise**.
- Other technologies like LSTM takes a lot more time for training and testing.

FUTURE PROSPECTS

- We can add the feature of analyzing real-time audio input using either audio recorder software or extracting audio from a real-time video feed.
- Secondly, to increase the accuracy, we can also make our model gender-sensitive, meaning a gender-based recognition model where gender classification is performed first, followed by an emotion classification model of each gender.

THANK YOU