NORTH SOUTH UNIVERSITY

DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING

CSE 499 Senior Design Thesis Report

Real-Time Emotion Recognition from Face Expressions Using Deep Learning Techniques

A Dissertation Submitted to the Department of Electrical and Computer Engineering of North South University

In the Partial Fulfilment of the Requirements for the Degree

Of

Bachelor of Science in Computer Science and Engineering

By

Erfan Mostafiz
ID: 1912734042
erfan.mostafiz@northsouth.edu

Shukdev Datta
ID: 1911838042
shukdev.datta@northsouth.edu

Md Toufique Husein
ID: 1921750642
toufique.husein@northsouth.edu

Under the Supervision of

Dr. Shafin Rahman
Assistant Professor
Department of Electrical & Computer Engineering
North South University
Dhaka, Bangladesh

Fall 2022

DECLARATION

It is hereby acknowledged that:

- No illegitimate procedure has been practiced during the preparation of this document.
- This document does not contain any previously published material without proper citation.
- This document represents our own accomplishment while being Undergraduate Students at **North South University**

Sincerely,

Erfan Mostafiz (ID: 1912734042)

Department of Electrical & Computer Engineering,
North South University

Shukdev Datta (ID: 1911838042)

Department of Electrical & Computer Engineering,
North South University

Md Toufique Husein (ID: 1921750642)

Department of Electrical & Computer Engineering,
North South University

Date: 5th January, 2023

APPROVAL

This is to certify that this senior design capstone project report entitled "**Real-Time Emotion Recognition from Face Expressions using Deep Learning Techniques**", submitted by Erfan Mostafiz (Student ID: 1912734042), Shukdev Datta (Student ID: 1911838042), and Md Toufique Husein (Student ID: 1921750642) are undergraduate students of the **Department of Electrical & Computer Engineering, North South University**. This report partially fulfils the requirements for the degree of Bachelor of Science in Computer Science and Engineering on January 2023, and has been accepted as satisfactory.

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation.

Supervisor
Dr. Shafin Rahman

Assistant Professor
Department of Electrical & Computer Engineering
North South University
Dhaka, Bangladesh

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation.

Dr. Rajesh Palit

Professor & Chair
Department of Electrical & Computer Engineering
North South University
Dhaka, Bangladesh

AKNOWLEDGEMENT

We hereby acknowledge the continuous guidance, support and encouragement we received from our honorable supervisor Dr. Shafin Rahman, towards the accomplishment of this project. We would like to express our profound gratitude to our honorable supervisor, Dr. Shafin Rahman, for his diligent guidance, valuable feedbacks, patience and encouragement towards the completion of this project. Without his support, it would not have been possible for us to finish our project.

We dedicate this piece of work to numerous people around the world who needs their emotions checked and get support and guidance of, especially in the post-covid19 era.

Finally, we would like to thank everybody who supported us and provided us with counsel for the completion of this project.

ABSTRACT

In today's fast paced world, detection and analysis of human emotion has become one of the major topics of research. This field of detecting human emotion through computational methods is called Affective Computing, an interdisciplinary field spanning computer science, psychology, and cognitive science. In this study we have proposed deep learning based models for emotion recognition using face expression image and EEG data. We have used two distinct state-of-the-art datasets in our work – the DEAP dataset and the FER2013 dataset. We have trained an LSTM network for EEG signals based emotion detection that uses EEG data from the DEAP dataset. For frontal face expression based emotion detection using the FER2013 dataset, we have used three deep learning architectures – Transfer Learning using MobileNet backbone, Convolutional Neural Network (CNN) and Attentional Convolutional Network. All three of these models use the FER2013 emotion dataset for training. The LSTM network for EEG based emotion gives an accuracy of 87% in the continuous valence-arousal-dominance scale of emotion. The MobileNet Transfer Learning model gave an accuracy of 95%, the CNN model an accuracy of 75.4% and the Attentional Convolutional Model gave an accuracy of 58% on the test data. Finally, we have designed a front-end interface with our best model which can predict emotions in real-time in live video feed.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Emotions are mental states caused by neurophysiological changes that are related with varied ideas, sensations, behavioral reactions, and a level of pleasure or dissatisfaction. It is one of the key components that affect people directly in their day-to-day lives. In today's world, especially in the post-covid era, detecting people's emotion has become an area of major interest. Emotions play an important role in how we think, behave and act in our daily lives. They are influenced by the limbic system, which is a network of interconnected structures in the brain. Emotions and behavioral reactions are heavily influenced by key components in the brain, such as the hypothalamus, hippocampus, amygdala, and the limbic cortex [1].

As humans, we try to interpret and understand other people's state of emotion with every interpersonal interaction we have with them, consciously or subconsciously, through facial expression, voice or body language. This is due to years of training and experience that our brains have on interpreting other people's emotion through visual or vocal cues. But there are a lot of added complexities when it comes to human emotion detection by a computer, mostly because they are not trained to recognize emotions and only understand mathematical 1s or 0s. As a result, there have been a lot of research on emotion detection by a computer using various modalities, especially in recent times. This research falls under the branch of Affective Computing, an interdisciplinary field comprising of a blend between computer science, psychology, and cognitive science [2]. It involves human-computer interaction in which a computer can interpret and appropriately respond to its human user's emotions and other stimuli.

1.1 Background and Motivation

Emotions play a very significant part in human life and is expressed by everyone in one way or other. As years pass by, Artificial Intelligence (AI) based systems are becoming highly intertwined with our lives. As a result, there has been a growing number of researches involved in making technologies that can understand and interact with humans, and a vast majority of them had been with detecting and analyzing human emotions. The fields in which emotion recognition has gained an increasing amount of interest include human-computer interaction [3] [4], animation [5], medicine [6] [7], augmented and virtual reality [8] [9], advanced driver assistance systems [10] and security systems [11].

There are different ways in which emotions can be detected which include facial expressions [12] [13] [14], speech [15], textual data [16], gaze direction [17] and bio-signals [18] including electroencephalogram (EEG) and electrocardiogram (ECG) data. Among these, facial expressions are one of the most prevalent ways of emotion detection. This is because in a human interaction, 7% of the emotional information is conveyed through spoken words, 38% by speech and tone, and the remaining 55% through facial expressions [19], thus leaving facial expression as the most important form in which human beings express emotions.

1.2 Study Objectives

In our research, with the help of deep learning technologies, we set out to predicting emotions from facial expressions. There are mainly two scales of detecting emotion – the continuous scale and the discrete or categorical scale. The continuous scale usually includes three dimensions – valence, arousal and dominance. All 3 scales range from 0 to 10. The valence scale denotes negative or positive emotion, with values close to 10 referring to more positive emotions while values close to 0 referring to more negative emotions. The arousal scale denotes calm or active state, with values close to 10 denoting more active state while values close to 0 denoting more calm state of mind. The dominance scale indicates if the person is dominated by an emotion or is in control, with higher values meaning more in-dominated by a particular emotion while lower values meaning more in-control of the emotion.

On the other hand, the discrete scale of emotions contains different categories of emotion in 2-dimensional space. According to Paul Ekman, an American psychologist and an expert in the study of facial expressions, there are seven basic categories of emotions – joy, surprise, sadness, anger, disgust, fear, and contempt [20]. The discrete categories of these universal emotions, according to Ekman, are explained below [21]-

1. Joy or happiness - This is indicated by raising of the mouth corners diagonally, cheeks raised and the eyelids tightening.
2. Surprise – This emotion is characterized by the bending of the eyebrows, eyes opened wide and exposing more sclera part of the eye (the white layer) and tightening of the eyelids.
3. Sadness – This emotion is symbolized by the lowering of the corners of the mouth, the descending of the eyebrows to their inner corner and the eyelids sagging downwards.
4. Anger – This emotion is illustrated by the lowering of eyebrows, the lips pressing tightly against the mouth and the eyes bulging.
5. Disgust – This emotion is often observed when a person raises their upper lip, their nose wrinkling, raising of the cheeks and eyebrows pulled down.
6. Fear – This emotion is symbolized by the raising of upper eyelids and the eyebrows, eyes opening wide and the lips stretching horizontally.
7. Contempt – This emotion is shown by the tightening up of half of the upper lip and often the head tilting slightly backwards.

In our research, we started out with detecting emotions in both categorical scale (2-dimensional) and continuous scale (multi-dimensional). For the detection of valence-arousal-dominance of the continuous emotion scale, we used the state-of-the-art DEAP dataset, which is a multimodal dataset consisting of EEG signals, video data and participant questionnaires of their emotions while watching music videos. For the detection of emotions in the categorical scale, we used two state-of-the art image datasets of emotions – the FER2013 and the CK+ dataset.

However, we could not train our emotion recognizer model from the video files of the DEAP dataset due to computational complexities and bugs which will be discussed with further details in a later section. Hence, we opted out to only implementing the emotion detection models in the categorical scale using the image datasets mentioned above.

1.3 Research Goals

The goals of this study are summarized as follows:

- Analyzing and collecting suitable datasets pertaining to emotions.
- Designing and implementing deep learning model architectures suitable for emotion detection through image and video datasets.
- Preprocessing those datasets according to the deep learning models in which they would be feed into.
- Splitting the datasets into train, validation and test sets.
- Training the models on the train data and validating the trained weights and biases over the validation data.
- Inference of the trained models on the test set and on random video and image files from the internet containing different emotions through facial expressions.
- Designing and developing a front-end interface for real-time emotion detection using the trained deep learning model with the highest accuracy.

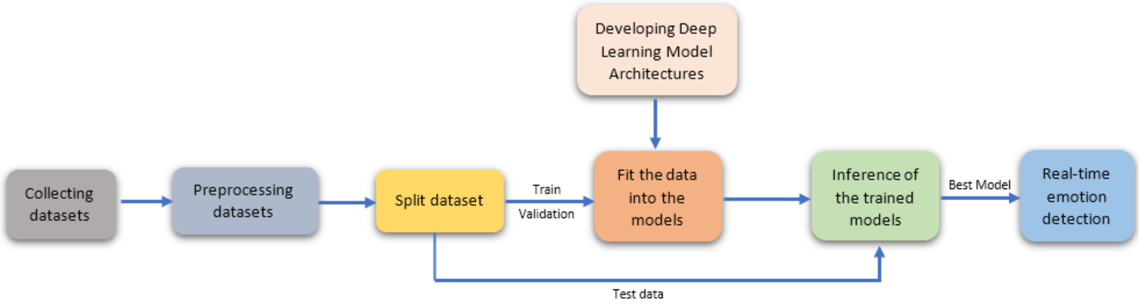

Developing Deep Learning Model Architectures
Collecting datasets
Preprocessing datasets
Split dataset
Train
Validation
Fit the data into the models
Inference of the trained models
Best Model
Real-time emotion detection
Test data

Figure 1.1- Research Goals

Our research goals are illustrated in Figure 1.1.

The rest of the paper is organized as follows: Chapter 1.3 talks about the research challenges we faced. Chapter 2 analyzes and reviews the related works in this field of study. Chapter 3 is dedicated to explaining the details of our project plan which includes the details of the datasets used and the preprocessing steps involved. Chapter 4 defines and explains all the deep learning models we have used for training the data. Chapter 5 talks about the experimental setup. Chapter 6 shows the results and findings of our work. Chapter 7 shows our real-time live emotion detection output. Chapter 8 talks about the impact of our work. Finally, the study has been concluded in Chapter 9.

1.4 Project Challenges

This section talks in detail about the challenges we faced during our research study and how we overcome those challenges.

1.4.1 Challenges with Video Based Emotion Detection

For detecting continuous scale of emotions in the valence-arousal-dominance space, we used the state-of-the-art DEAP dataset [22], which is a multimodal dataset consisting of EEG data and frontal face video. The participants' emotions were recorded as they each watched 40 one-minute-long portions of music videos. We designed two pipelines for continuous scale emotion detection – i) Emotion Detection using EEG signals, ii) Emotion Detection via Facial Action Coding System and Deep Learning techniques using frontal face videos. Although we could successfully implement our 1st pipeline using EEG signals, which is discussed in further details in a later section, we could not be successful in implementing our 2nd pipeline which uses frontal face videos.

In our 2nd pipeline which uses the frontal face videos of the DEAP dataset, we implemented the Facial Action Coding System (FACS) which gathers 3D landmark coordinates from the face. For this, we used Google's MediaPipe FaceMesh Library to collect 478 3D landmarks coordinates off the face of the participant videos, which includes 468 face landmarks and 10 additional iris landmarks of both the eyes (5 for each eye). The landmark coordinates and the corresponding frames were preprocessed. The "participant questionnaire" csv file of the DEAP dataset contained self-rated valence, arousal and dominance values of the participants. We used these emotion values as our ground truth. We developed two deep learning based models for video based emotion detection – a Long Short-Term Memory (LSTM) network, and a Video Vision Transformer network (ViViT).

But the problem came when we fed our input landmarks and the frames, and the output ground truth of the valence-arousal-dominance values into the two networks. Computational complexity and bugs occurred while fitting our training data into both of the networks. We could not get a fix of these bugs, and thus later shifted our focus to image based emotion detection using facial expressions.

Chapter 2: Related Works

There have been several works that have tried to detect emotions from different modalities (image, video, EEG, etc). This chapter reviews the related literature in this field.

2.1 Deep Neural Networks for Image based emotion classification

Inigo et al. [23] have proposed a custom CNN based architecture that can detect whether a person wears a facemask or not and the proposed CNN architecture contained four convolution layer, one fully connected layer and finally the output layer. After each Convolution layer, ReLU is used as activation function, which returns 0 if it receives any negative input but for positive input it returns that value of x. The ReLU here actually dropped the neurons having negative values but neurons having positive values are kept unchanged. They have also used two max pooling layers that caused dimension reduction. In order to avoid overfitting, they have used batch normalization and dropout. Softmax classifier is added after the last fully connected layer to produce the output. The output that is resulted from the first convolution layer is the input to the next convolution layer. The second convolution layer contained 64 3x3 filters and the third convolution layer contained 128 3x3x3 filters. The fully connected layers contained 128 neurons in total. Since the image dataset, Ck+ and JAFFE have greyscale image, so there is a need to identify the optimal weights and bias of the network which seemed to be a difficult issue. So they have divided each of the pixel values by 255 and then map those values in between -1 to 1. OpenCV is a python library that was used to detect the faces in the input image frames. The result showed that Happy face had an accuracy of 79.7%, Sad face had an accuracy of 39.2% and Angry face had an accuracy of 56.4%. The model accuracy graph showed the proposed model had a test accuracy of 67% and training accuracy of 66%.

Siam et al. [24] have proposed a technique that consisted of 4 phases, which are image processing carried out by SRGAN followed by MediaPipe, which will create landmarks on the face. Then a key landmark analysis phase. This will produce key landmark position in face and generate the mesh angular encoding. Then the feature map produced from the mesh angular encoding will be fed into the Classification phase that contains the ML models like DT, SVM, KNN, RF, QDA and MLP. The classifiers will predict one of the seven selected categories of emotion. Facial Action Coding system is introduced here which will take account of the movement of facial muscles when expressing emotions and out of 468 landmarks generated by MediaPipe, only 27 key landmarks are taken. They have chosen three datasets to work on which are CK+, JAFFE, and RAF-DB. KNN model had the highest accuracy of 97% in CK+ dataset and Gaussian NB had the least accuracy of 84% in CK+ dataset. KNN model had the highest accuracy of 95% in JAFFE dataset and QDA had the least accuracy of 79% in JAFFE dataset. MLP and SVM both had the highest accuracy of 67% and DT had the lowest accuracy of 53% in RAF-DB dataset.

Kaviya et al. [25] have proposed a method that used custom dataset and FER2013 dataset. During processing phase, they have converted the RGB image to grey scale image. They have used Haar filter from OpenCV, which will act as the face detector and detect face from static and dynamic image. Then the images are resized to 48x48 and facial features are detected along with facial landmarks. The CNN proposed architecture will take landmarks in the form of input. The output produced from the CNN is fed into the Speech Synthesizer that will produce output in the form of audio. The proposed model gave test accuracy of 65% in FER-2013 and 60% accuracy on Custom Dataset.

Lasri et al. [26] have proposed a method that will take face image from students and the face image is cropped to a dimension of 48x48 and then normalized. Then the image is fed into the CNN in the form of input. The convolution layer in CNN will be used as a feature extractor to extract feature from input image. Then pooling layer is used which will reduce the dimensionality of the feature maps generated in the earlier stage and we have done max pool which takes max from each block. Then fully connected layer is used as the form of output phase that will produce the outputs. The happy emotion detected from the face image has a precision value of 0.88, Recall value of 0.90 and F1-score of 0.89, which is max among all the seven categories of emotions. They have run their CNN model for 106 epochs and they have acquired an accuracy of 70%.

2.2 Deep Neural Networks for Video based emotion classification

Verma et al. [27] have proposed a 3D model that is concerned with three continuous emotions, which are valence, arousal and dominance. They have used DEAP dataset that gave them the data for the valence, arousal, and dominance rating for large number of emotions. It is very important to understand the location of all the discrete emotions (fun, happy, joy, cheerful, melancholy, depressing, terrible, exiting, love, lovely, sentimental, sad, mellow, shock, hate etc) in the Valence-Arousal-Dominance Space Diagram. They have conducted 40 trials for each of the 32 participants. The researchers have plotted 1280 points in the VAD space and the mean and standard deviation of all the discrete emotions are calculated on VAD dimensions. For example, in Valence space of VAD diagram, mean of Fun is found to be 6.8571 and SD of Fun is 1.3015. In the same way, mean of Fun and SD of Fun are calculated in Arousal and Dominance Space in VAD diagram. The researchers have performed K-means clustering which resulted in 5 distinct clusters which are C1, C2, C3, C4 and C5. In addition, they have calculated the Euclidean distance between the discrete emotions. C1 cluster contained discrete emotion like Happy, Joy, Fun, Exciting and Cheerful, C2 cluster contained discrete emotion like Love and Lovely, C3 cluster contained discrete emotion like Sentimental and Mellow, C4 cluster contained discrete emotion like Sad, Depressing and Melancholy, and C5 cluster contained discrete emotion like Shock, Hate and Terrible. They have used three types of data, which are Video, EEG signals, and VAD data in VAD space diagram. After preprocessing of these data types, they are sent to the emotion classification phase, which contained three ML models, which are MLP, SVM and K-NN. MLP scored 63.47% classification rate for valence in EEG dataset and 98.43% for valence in Video dataset. MLP scored 69.62% classification rate for arousal in EEG dataset and 99.43% for arousal in Video dataset. MLP scored 63.57% classification rate for dominance in EEG dataset and 98.53% for dominance in Video dataset. SVM scored 56.34% classification rate for valence in EEG dataset and 96.20% for valence in Video dataset. SVM scored 52.79% classification rate for arousal in EEG dataset and 99.63% for arousal in Video dataset. SVM scored 57.71% classification rate for dominance in EEG dataset and 97.06% for dominance in Video dataset. KNN scored 67.51% classification rate for valence in EEG dataset and 98.46% for valence in Video dataset. KNN scored 68.55% classification rate for arousal in EEG dataset and 99.30% for arousal in Video dataset. KNN scored 65.10% classification rate for dominance in EEG dataset and 98.36% for dominance in Video dataset.

XIANZHANG et al. [28] have proposed an end-to-end model for VFER which was greatly influenced by spatial(space) and temporal(time) information. They have designed two streams which are spatial stream and temporal stream. The spatial stream is used to analyze the facial structure of a face and temporal stream is used to analyze the movement of the facial muscles. Both the streams used CNN and LSTM. CNN can only process a single image and has a drawback since it cannot directly process more than one image. CNN is basically used to extract essential features from the image. LSTM can process dynamic vectors but it has a drawback also which is that it cannot process matrices. So it has been used to process the sequence vectors generater by the CNN. Then finally they have used a layer named “Aggregation Layer” that is used to extract the output of the temporal and spatial streams. The temporal stream is designed to understand the importance of facial muscles movement when expressing emotion and spatial stream is designed to understand the importance of spatial information like texture of face when expressing emotion. The datasets they have used are RML and eNTERFACE05 dataset. The results have shown that their proposed model have outperformed previous related works and the accuracy scored on RML dataset was 65.72% and accuracy scored on eNTERFACE05 was 42.98%.

Zheng et al. [29] conducted research on emotion identification based on physiological cues. They have recently established an accessible EEG-based emotion identification database. The binary (happy vs. calm) classification rate for the emotion-related database created by additional researchers, which consists of four physiological signals from the ECG, galvanic skin response, skin temperature, and respiration, achieved 86.7% Accuracy.

Trainor et al. [30] employed music to elicit four different emotions. They discovered that while utilizing positive musical materials, the frontal areas of the left hemisphere experienced heightened EEG activity, and when using negative musical materials, the frontal portions of the right hemisphere experienced enhanced EEG activity. The authors conclude that the frontal lobes of the human brain and emotion are closely related. They obtained 67.7% classification accuracy.

Zhang et al. [31] proposed a method that consisted of three pipelines in their proposed architecture. First one is face appearance pipeline that contains CNN of 3 layers and that CNN is used to pretrain the AFEWVA dataset. It is to be done so that it can capture detailed spatial information. The input image of this pipeline has a dimension of 64x64 size. The first layer had 32, second one 64 and third one had 128 filters. These filters had a dimension of 3x3. All the three layers had ReLu and 3x3 max pooling feature included. The ConvLSTM layer helps to extract features from image with respect to space and time. Next pipeline is the bio-sensing data pipeline. In this pipeline, the researchers have used 3 1D convolution layers along with a fully connected layer. The 1D convolution layer all had ReLU and batch normalization layer added to them and the FC layer is used for feature extraction purpose and for the process of flattening. The bio signals are taken via EEG and ECG and GSR data. The researchers have used required equipment to measure these signals. Classification Pipeline is the final pipeline of their proposed architecture. Here the output from the first and second pipeline merged to create multi-modal feature vector. In case of classification, Adaptive Boosting is used which predicts the fusion score of all the sub-classifier that does prediction based on the results obtained from the facial and bio signal features. Adaptive Boosting decreases the weights on the sub-classifier if the prediction is correct or vice versa. The performance based of DEAP dataset was 98.56 percent/98.18 percent.

Chapter 3: Project Plan

Starting from 'Deciding work with Emotion Analysis' and collecting the Dataset from Deap data, preparing them for converting video to image dataset, several intermediate videos were corrupted. Approaching the tasks with an earlier plan. And then, we create the LSTM model to know how accuracy comes, and finally, we choose another dataset, FER2013. Using this dataset, prepare the data for Attention CNN, CNN. Also, we use the MobileNet model to get better accuracy. The project plan is shown in the figure below (Figure 3.1).

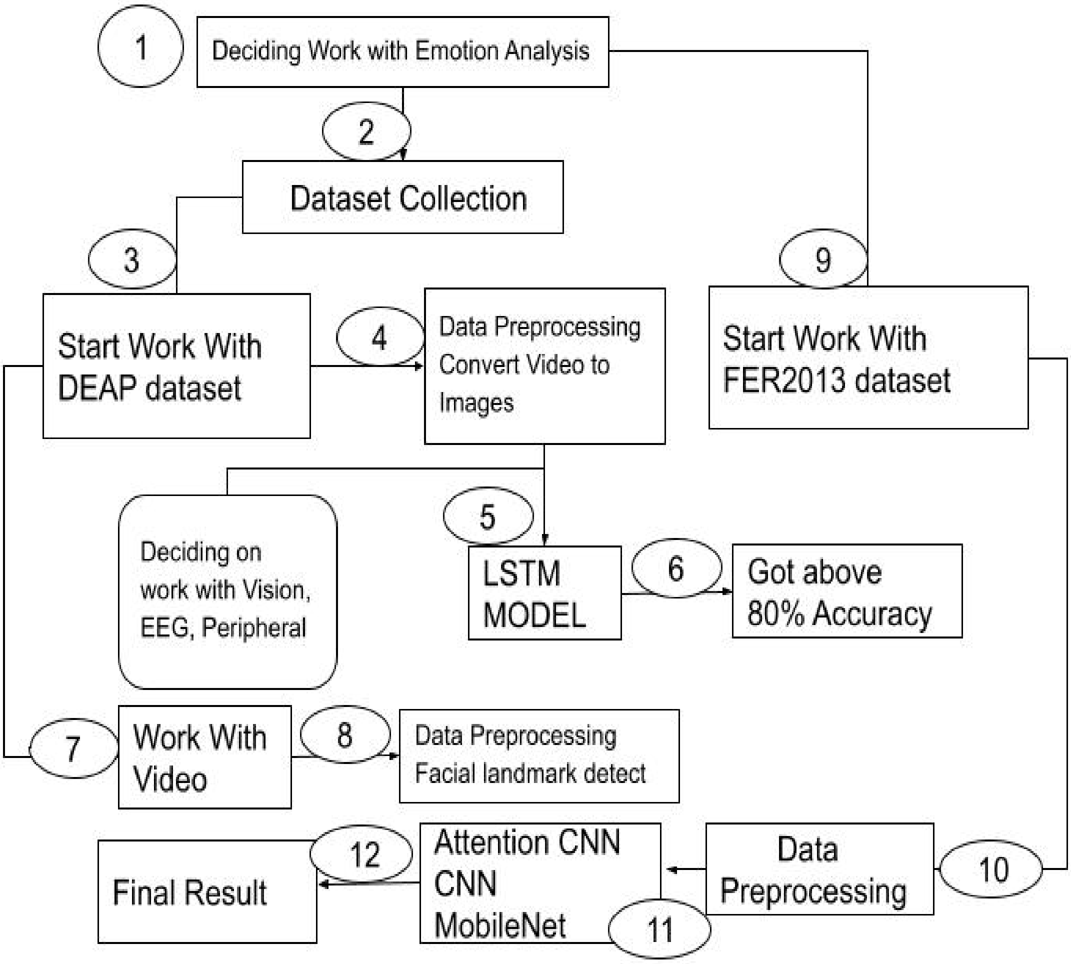

1
Deciding Work with Emotion Analysis
2
Dataset Collection
3
Start Work With DEAP dataset
4
Data Preprocessing Convert Video to Images
9
Start Work With FER2013 dataset
Deciding on work with Vision, EEG, Peripheral
5
LSTM MODEL
6
Got above 80% Accuracy
7
Work With Video
8
Data Preprocessing Facial landmark detect
Final Result
12
Attention CNN CNN MobileNet
11
Data Preprocessing
10

Figure 3.1 - Project Plan. Numbers in the Figure Denotes the Chronology of the Tasks

The chronology of the tasks during our study is indicated by the numbers in the figure. The Final Result will be deployed via a Website or an Android App.

3.1 Description of the Plan

Initially, we chose a path that works with the DEAP dataset. In this dataset, the results of an online self-evaluation in which 32 volunteers scored each of 120 one-minute music video excerpts based on arousal, valence, and dominance. However, after a few months of working with that, we devised another plan—the latest work done with the FER2013 dataset. Using the Attentional convolutional neural network, convolutional neural network, and MobileNet, we predict a person's emotion by reading their facial expression.

Five layers of a convolutional neural network have been proposed by the researchers. The FER2013 dataset, comprised of grayscale photographs of various emotions in 48x48 pixel sizes, was employed in their research. There were 28709 photos in total in this dataset. The first layer will be an input layer that receives a black-and-white image with a dimension of 48x48 pixels. A 5x5 filter with 64 filters and 100 feature extractions is coupled to the first CONV layer. The essay provides every last architectural detail. After CONV2D layers, max pooling is used to minimize the number of pixels in the image, hence lowering the image's dimension. Max pooling reduces the parameters and the image resolution, which lowers the computing effort. To convert a two-dimensional to one-dimensional input, utilize the flattened layer. Neuronal regularization uses dropout. Relu was utilized as the activation function in the first two dense layers. At the same time, Softmax was employed in the last dense layer, which divided the output into seven groups.

In the last phase, when 25 epochs are utilized, the model produced a test accuracy of 57.481193 percent and a test loss of 3.743610 percent. The model also estimates a 0.429 percent train loss and a 98.766 percent train correctness. The model performed exceptionally well for positive emotions—78% precision for pleased and 75% for a surprise—but poorly for negative emotions.

3.2 Dataset Description

In our work, we aspired to help people identify their emotions through object detection and face recognition. Two datasets have been used in our project – the DEAP dataset and the FER2013 dataset.

3.2.1 DEAP Dataset

The DEAP dataset details is given in Table 3.1. The EEG signals of 32 participants were recorded as each watched 40 1-minute-long portions of music videos. Participants rated each video in terms of the levels of arousal, valence, dominance, like/dislike and familiarity. The DEAP dataset has been used for EEG based Emotion Detection.

Dataset link: https://www.eecs.qmul.ac.uk/mmv/datasets/deap/

Each of these 32 preprocessed.dat files collected from the DEAP dataset page contained two arrays: data and labels, one for each of the 32 participants. Data has dimensions of 40,40,8064. Each video had 40 channels and had 8064 EEG signal data points, for a total of 322560. The labels were in the shape of a 40 by 4, where 4 stands for valence, arousal, dominance, and liking. NumPy arrays are loaded and used in Python.

Criterion	DEAP Dataset
Dataset Size	22 subjects x 40 trials
Stimulus duration	60s
Modality	Vision, EEG, Peripheral
Used bio signals	EEG, EOG, EMG, GSR, Respiration belt, Plethysmograph, Temperature
Labels	Continuous valence and arousal from 1 to 9

Table 3.1 - Details of the DEAP Dataset

3.2.2 FER2013 Dataset

The FER2013 dataset, on the other hand, contains around 35K facial Grayscale images of 7 types of emotions, and all the image dimensions were 48x48 pixels. 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. Details of the FER2013 dataset is given in Table 3.2. Sample images in the FER2013 dataset is given in Figure 3.2.

Criterion	FER2013 Dataset
Dataset Size	35,887 facial images
Image dimensions	48 x 48 pixels
Image Type	Grey Scale Images
Emotion Categories	7 classes of emotions
Labels	0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral

Table 3.2 - Details of the FER2013 Dataset

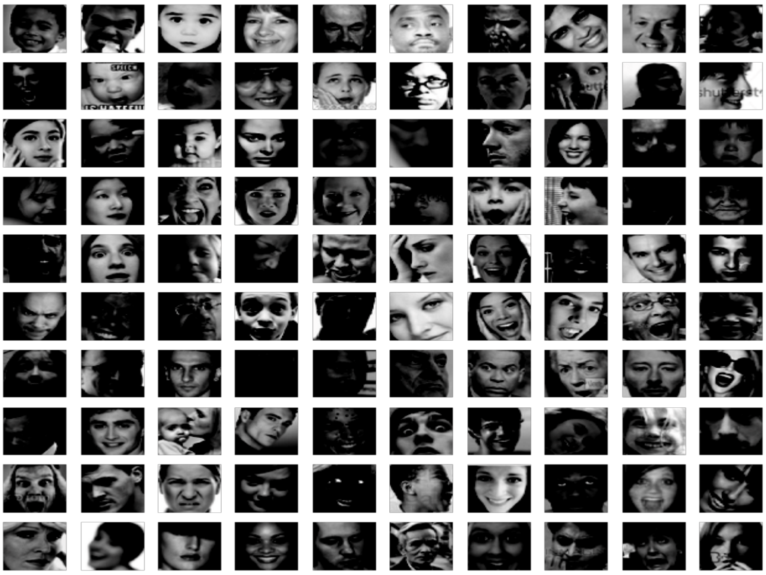


Figure 3.2 - FER2013 image samples

3.3 Data Preprocessing

We used the FER2013 dataset. The data collection is made up of grayscale portraits of 48x48-pixel faces. Each face occupies the same amount of space in each image and is roughly in the middle, thanks to the automatic registration of the faces.

Each face must be categorized into one of seven categories based on the emotion each face represents (0 = furious, 1 = disgust, 2 = fear, 3 = pleased, 4 = sad, 5 = astonished, 6 = neutral). The training set has 28,709 samples, whereas the validation set contains 3,589 samples, and the test set contained 3589 images.

The dataset was not initially balanced, as shown in Figure 3.3. There were very few instances of the "Disgust" class of emotion compared to "Happiness" class of emotion which was too many. Thus, the dataset needed to be balanced, otherwise the trained model would be heavily biased towards one or two class.

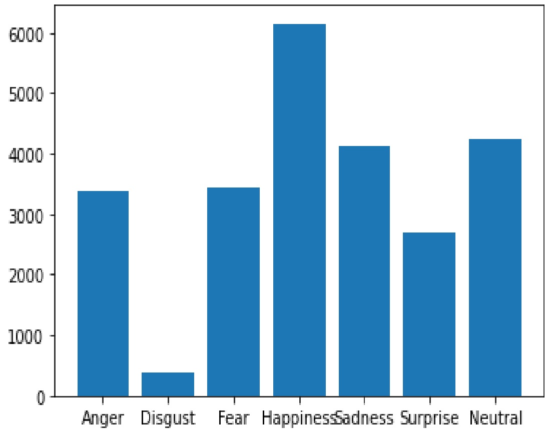


Figure 3.3 - FER2013 Before Balancing

After balancing, the dataset now had approximately the same number of images for each class. The balanced dataset is shown in Figure 3.4. After balancing the dataset, the dataset was reduced to 29,023 total images, which was still more than enough for our model training.

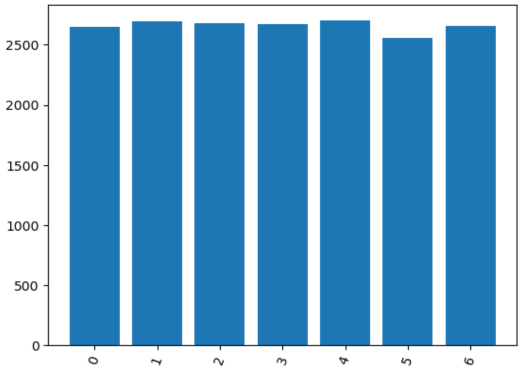


Figure 3.4 - FER2013 After Balancing

The balanced dataset was then split into 80:20:20 ratio for train:validation:test data respectively. The dataset splitting part is illustrated in Figure 3.5.

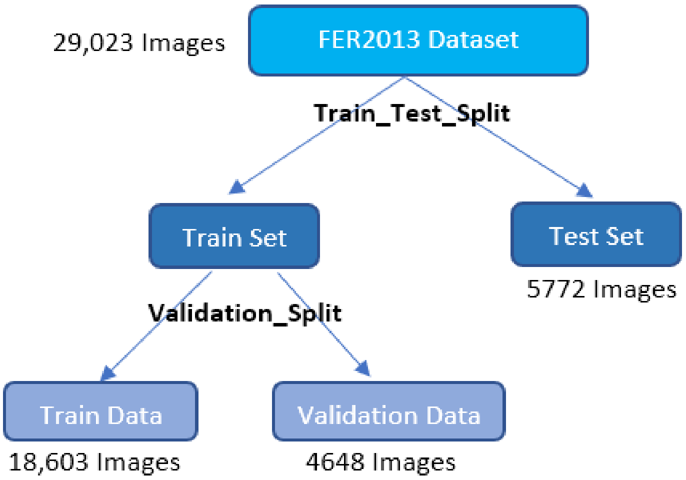


Figure 3.5 - Splitting FER2013 dataset

Chapter 4: Methodology

4.1 Transfer Learning using MobileNet Architecture

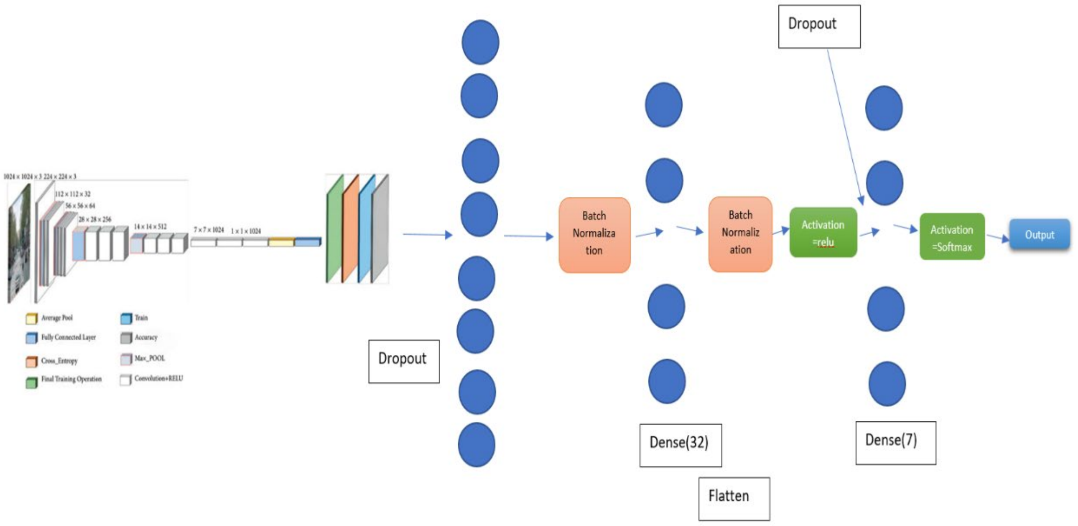


Figure 4.1 - Transfer Learning using MobileNet Architecture Diagram

MobileNet [32] has introduced a new way of convolutional layer, known as Depthwise Separable convolution [33]. MobileNet is very small network that makes it very compatible using in mobile or any other embedded devices. MobileNet requires very small space in Disk to load which is around 16-18MB. Depthwise convolution is different from ordinary CONV2D because CONV2D performs convolution on all the input channels but in case of Depthwise convolution, each input channel kept separate.

A closer look into the MobileNet Architecture is given in Table 4.1. One thing to notice about this architecture diagram above is that the input first and second dimension is being halved each time convolution is performed because of having stride of 2 and padding of 0. Detailed difference between standard convolution layer and depthwise separable convolution layer in MobileNet is given in Figure 4.2.

Type/Stride	Shape of filter	Size of Input
Convolution layer/stride 2	3x3x3x32	224x224x3
Convolution depthwise/stride 1	3x3x32 depthwise	112x112x32
Convolution/stride 1	1x1x32x64	112x112x32
Convolution depthwise/stride 2	3x3x64 depthwise	112x112x64
Convolution/stride 1	1x1x64x128	56x56x64
Convolution depthwise/stride 1	3x3x128 depthwise	56x56x128
Convolution/stride 1	1x1x128x128	56x56x128
Convolution depthwise/stride 2	3x3x128 depthwise	56x56x128
Convolution/stride 1	1x1x128x256	28x28x128
Convolution depthwise/stride 1	3x3x256 depthwise	28x28x256
Convolution/stride 1	1x1x256x256	28x28x256
Convolution depthwise/stride 2	3x3x256 depthwise	28x28x256
Convolution/stride 1	1x1x256x512	14x14 x256
5x Convolution depthwise/stride 1, Convolution/stride 1	3x3x512 depthwise, 1x1x512x512	14x14x512, 14x14x512
Convolution depthwise/stride 2	3x3x512 depthwise	14x14x512
Convolution/stride 1	1x1x512x1024	7x7x512
Convolution depthwise/stride 2	3x3x1024 depthwise	7x7x1024
Convolution/stride 1	1x1x1024x1024	7x7x1024
Average Pooling/Stride 1	7x7 Pool	7x7x1024
Fully connected/Stride 1	1024x1000	1x1x1024
Softmax/Stride 1	Classifier	1x1x1000

Table 4.1 -MobileNet backbone details *[32]*

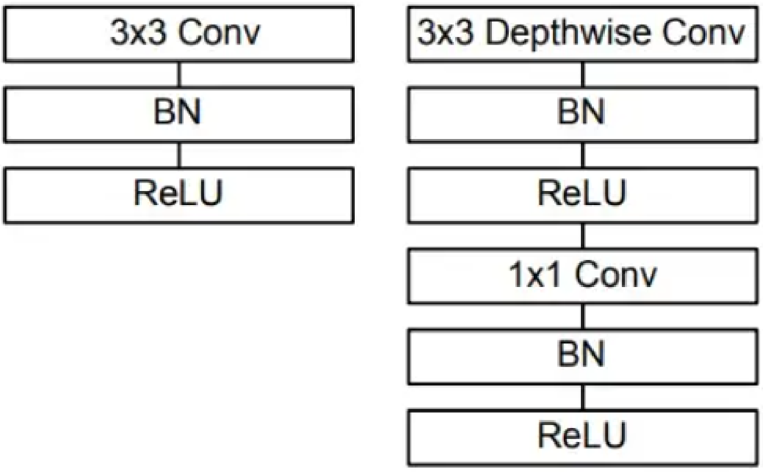


Figure 4.2 - Left: Standard Convolutional layer, Right: Depthwise Separable Convolutional layers in MobileNet *[32]*

The diagram above shows that the normal typical convolution contains 3x3 Conv followed by BN and then ReLU as activation function.

However, in case of Depthwise Separable Convolution Layer, we have 3x3 Depthwise Conv followed by BN, which is later followed by ReLU in the form of activation function. However, there is more to it, which is 1x1 Conv, which act as the bottleneck convolution that causes the dimensionality reduction. Next again BN is there and followed by ReLU as activation function.

The whole procedure of Depthwise Separable Convolution is discussed below in Figure 4.3.

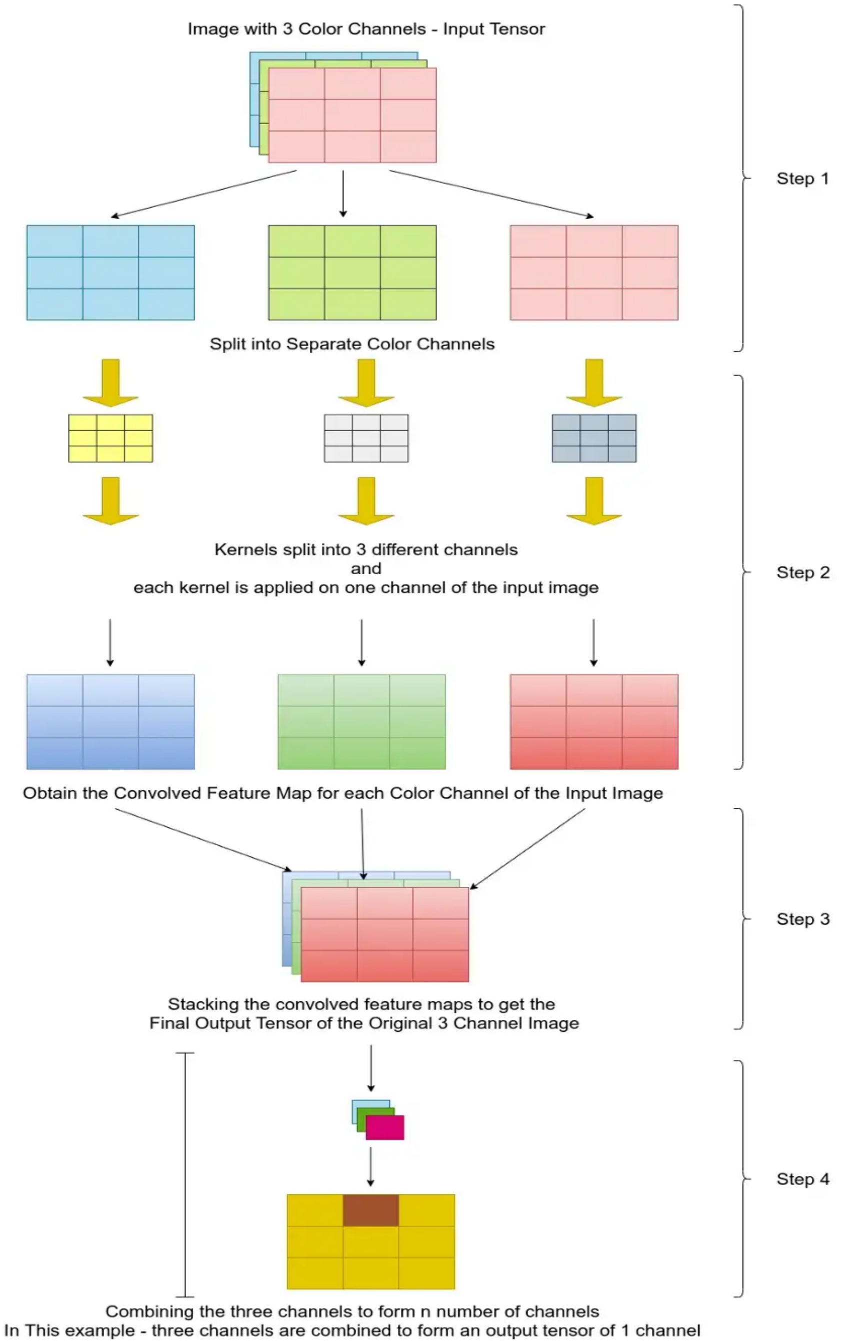


Image with 3 Color Channels - Input Tensor
Step 1
Split into Separate Color Channels
Kernels split into 3 different channels
and
each kernel is applied on one channel of the input image
Step 2
Obtain the Convolved Feature Map for each Color Channel of the Input Image
Stacking the convolved feature maps to get the
Final Output Tensor of the Original 3 Channel Image
Step 3
Step 4
Combining the three channels to form n number of channels
In This example - three channels are combined to form an output tensor of 1 channel

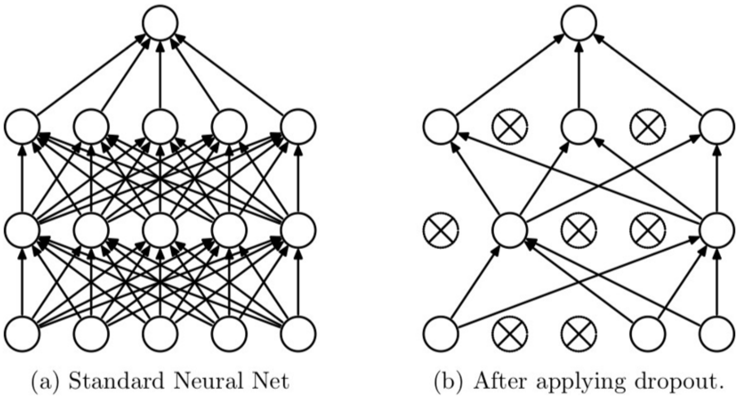
Figure 4.3 - Explanation of Depthwise Separable Convolutions *[32]*

After the step 4, the output of the Depthwise Separable convolution fed to the first dropout layer and then the dropout layer is flattened whose output shape is (None, 1024). After flattening, batch normalization is performed which zero centers the input by calculating the mean and standard deviation and then variance. Normalization equation is performed to normalize the inputs and the inputs are rescaled again for more optimization to get better output. Then the output is passed to the first dense layer where the activation is ReLU that produces output of 1 if the value is positive or 0 if the value is negative. Batch normalization is again performed. Output is again passed on to the next dropout layer followed by Dense layer that again comprises of Batch Normalization with activation as ReLU. This same sequence of Dropout, Dense, Batch Normalization and ReLU activation is followed again until final the output is passed on to the last Dense layer that contains Softmax as activation which produces the final output which is one of the seven discrete emotions [33].

The preprocessing phase and data preparation phase is covered in the "Data Preparation" section.

Hyper-parameter used in this combined architecture are 100 Epochs, 64 Batch Size, Categorical Loss Entropy as Loss function, Adam optimizer, Accuracy as the performance metrics, and 1e-3 as learning rate.

Use of Dropout in this architecture:


(a) Standard Neural Net

(b) After applying dropout.

Figure 4.4 - Use of Dropout

Dropout adaptively prunes the network. It is very inexpensive but very powerful tool. We turn off some of the neurons in the network-using dropout. Some of the forward pass paths are cut off. Therefore, this will make the network smaller. Dropout does not rely on single weight value too much. Weights are distributed in the whole network evenly. As mini-batch gradient descent is performed, the dropout layer can choose different subset at a time, and number of subsets is exponential, so it can train exponential number of different networks at a time. Dropout scheme is so intelligent that it does not need to do any voting. It forces the model to learn although if there is missing input or hidden unit.

4.2 Convolutional Neural Network Architecture

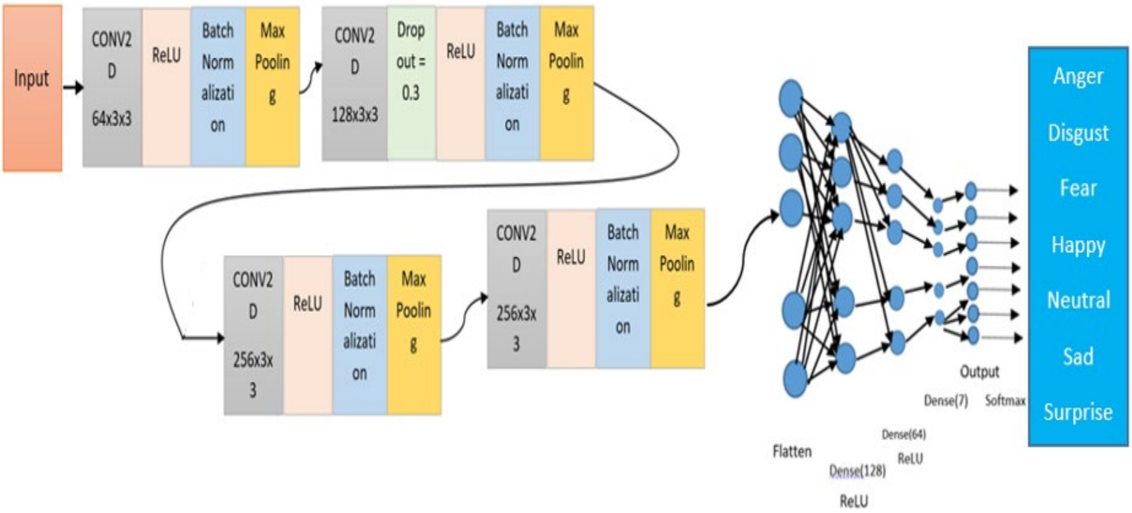


Figure 4.5 - CNN Architecture used

The proposed CNN architecture diagram is given above. The architecture has 4 CONV2D layer, 1 flatten layer, and 3 dense layers. Image is taken as input to the first CONV2D layer that has ReLU as activation function. ReLU activation function will give output of 1 if the value is positive but output of 0 if the value is negative. The input is zero centered using batch normalization. In order to zero center the inputs, the mean and standard deviation is calculated along with variance. Then the batch normalization equation is used to normalize the inputs. The inputs are rescaled for more optimization to get better outputs. Max pooling is applied that takes the maximum of the region where the filter is placed for convolution. The first CONV2D layer hase 64 filters of 3x3 dimension. After the operation in first CONV2D layer is finished, the output is sent to the next CONV2D layer in the form of input. The second CONV2D layer has 128 filters of 3x3 dimension. Dropout that adaptively prunes the network follows it. It is very inexpensive but very powerful tool. We turn off some of the neurons in the network-using dropout. Some of the forward pass paths are cut off. Therefore, this will make the network smaller. Dropout does not rely on single weight value too much. Weights are distributed in the whole network evenly. As mini-batch gradient descent is performed, the dropout layer can choose different subset at a time, and number of subsets is exponential, so it can train exponential number of different networks at a time. Dropout scheme is so intelligent that it does not need to do any voting. It forces the model to learn although if there is missing input or hidden unit. After Dropout layer, there is ReLU in the form of activation function and then used batch normalization layer and max pooling layer next to it. The third CONV2D is same as the first CONV2D layer because it follows the same sequence of first CONV layer. The only difference is that the convolution layer has 256 filters of 3x3 dimension. The fourth CONV2D layer is the same as the third one. Fourth convolution layer output is passed on to the next layer, where the input is flattened in order to convert the 2D arrays to single long continuous linear vector. After flatten layer, we have three Dense layers among which first two of them have activation of ReLU and the third one has activation of softmax. The first dense layer also contains dropout layer with it. The softmax will give the resultant output in the form of probability within 0 and 1. It never gives output of direct 0 or 1.

Hyper-parameter used in this CNN architecture are 15 Epochs, 64 Batch Size, Categorical Loss Entropy Loss function, Adam optimizer, Accuracy performance metrics.

4.3 Attentional Convolution Neural Network Architecture

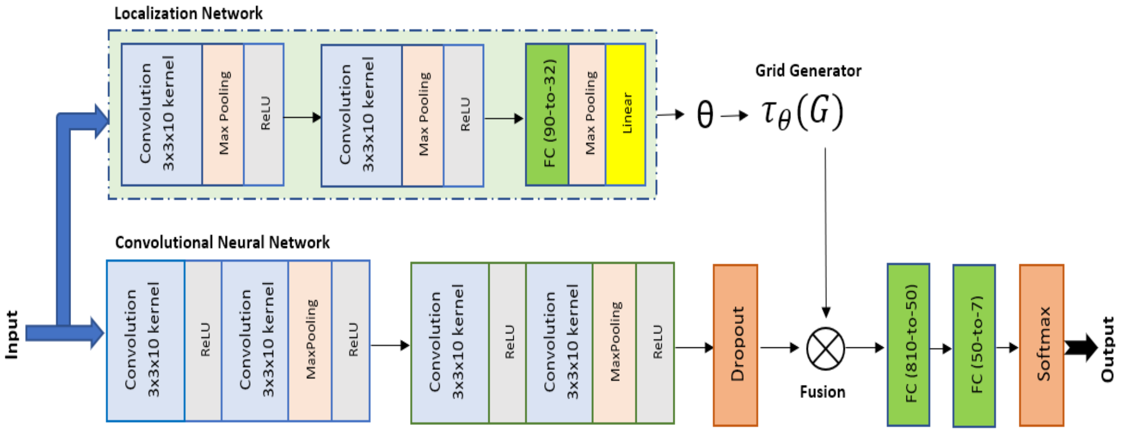


Figure 4.6 - Attention Based Convolution Network Architecture

The end-to-end attentional convolution network developed for our study is illustrated in Figure 4.6. This architecture demonstrates the idea of using a small convolutional neural network with less than 10 layers, and an attention/localization network which is trained from scratch. This is because for facial emotion recognition, the number of classes is small, and therefore, does not require a much larger network.

The idea behind this architecture is that, in a particular face expression image, only a prominent region of the face, from which most of the emotion is coming, needs to be given more attention, as not all parts of the face give out useful information for detecting emotion. Hence, a localization network is added to our framework which uses spatial information to highly focus on important face regions rather than spending computational time and space on other less important regions [34].

Given an input image containing facial expression, it will be fed into both the convolutional network and also the localization network. The convolutional network is also the feature extraction part which extracts features from the image. There are four convolutional layers in this network, each of kernel size 3x3 and 10 output channels. Each CNN layer is followed by a Rectified Linear Unit (ReLU) activation function. The ReLU is a function in which if the input is positive, then it becomes the output, but if the input is negative then a zero becomes the output [35].

$$ReLU\ f(x) = \max{(0, x)} \tag{4.1}$$

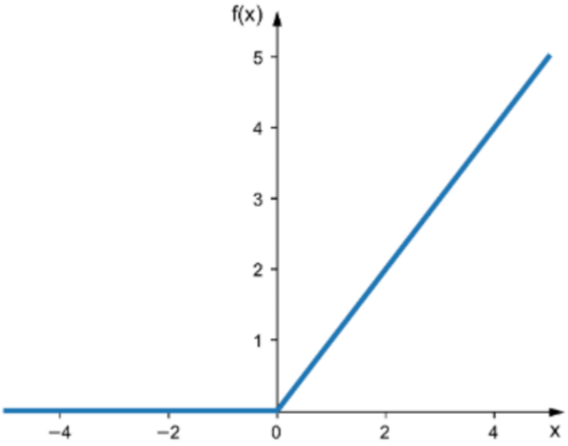
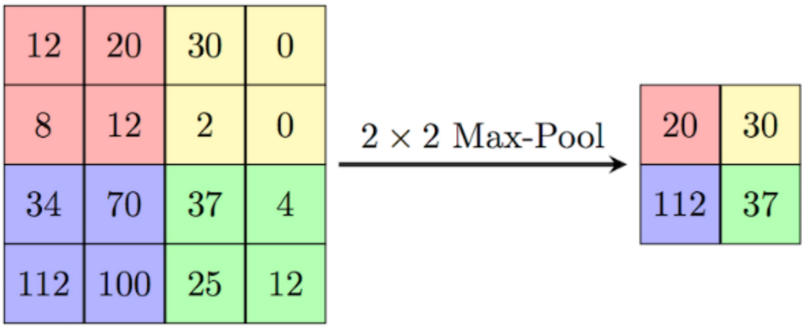


Figure 4.7 - ReLU Activation Function

After the 2nd and the 4th Convolutional layer, there is a MaxPooling layer. It is an operation that determines the maximum value in each portion of a feature map. The pooled feature maps highlight the most important features in the patch [36]. More specifically, MaxPooling2D has been used with kernel_size=2, stride=2, padding=0 and dilation=1. Figure 4.8 illustrates the use of MaxPooling. A dropout layer is then followed by.



12	20	30	0
8	12	2	0
34	70	37	4
112	100	25	12

20	30
112	37

Figure 4.8 - MaxPooling2D

On the other hand, the localization attention network consists of two convolutional layers, the first one with shape 8x3x3 and the second one with shape 10x3x3. Both layers are each followed by MaxPooling2D and ReLU activation function. The outputs are then passed onto a Fully Connected layer followed by ReLU and Linear Function. The regressed parameters are then transformed (using an affine transformation) into a sampling grid $T(\theta)$ producing the warped data. It is then fused with the data from the convolutional feature extraction layer, after which there are two fully connected layers and finally a softmax activation function. The SoftMax function finds the probability of each class according to the proportion of the values in the vector.

$$SoftMax = s(x_i) = \frac{e^{x_i}}{\sum_{j=i}^{n} e^{x_j}} \tag{4.2}$$

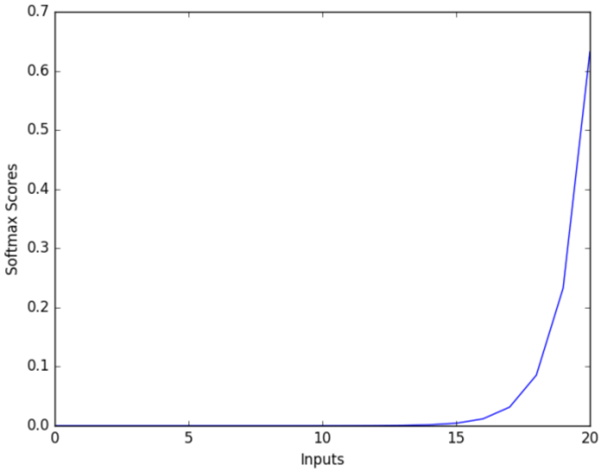


Figure 4.9- SoftMax Activation Graph

4.4 LSTM Network for EEG based Emotion Detection

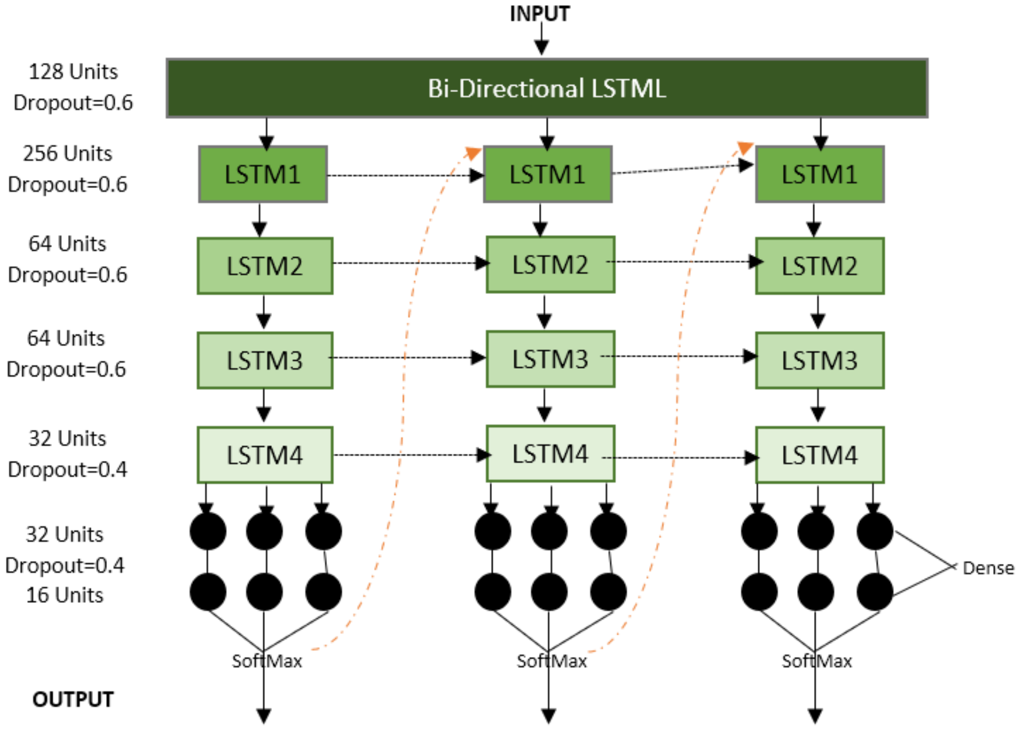


Figure 4.10 - LSTM Architecture for EEG based emotion detection

The Long Short-Term Memory (LSTM) architecture that we have used for EEG signal-based emotion detection is illustrated in Figure 4.10. In this network, 1 bi-directional LSTM layer, 4 LSTM layers and 2 dense layers have been used. The first bi-directional LSTM layer has 128 units and in total 256 units (128*2) and is followed by a dropout layer of 0.6 probability. The next layer is again a 256 neuron LSTM layer, followed by dropout layer of 0.6. The next two LSTM layers consist of 64 neurons and dropout layer. The final LSTM layer has 32 neurons with a dropout of 0.4. Then a dense layer of 16 units is used with ReLU as the activation function. Followed by this, another dense layer of 10 classes is used because our final output should be multiclass probability distribution over our 10 classes of emotion taken from the DEAP dataset.

Chapter 5: Experiment Setup

5.1 Hardware Specifications

The complexity of the models and datasets being utilized and the necessity for shorter training durations and more excellent performance motivate the development of better deep learning hardware. It is feasible to get good outcomes and apply algorithms for deep learning more effectively by using more powerful hardware. Table 5.1 displays the details of our hardware.

Component	Description
CPU	Intel Core i5-11400H, 11th generation
CPU Speed	4.5 GHz
RAM	8GB DDR4
GPU 1	Nvidia GeForce RTX 3050
GPU 1 Memory	4 GB
GPU 2	Intel® UHD Graphics
GPU 2 Memory	4 GB

Table 5.1- Hardware Specifications for this project

5.2 Software Specifications

Advanced python software and deep learning libraries had to be used to implement this project.

Software/Coding language and libraries	Description/ Version
Operating System	Windows 10 Pro
Python	Version 3.10
Pip	Version 22.3
TensorFlow	Version 2.11
Keras	Version 2.11
PyTorch	Version 1.12
Anaconda Jupyter Notebook	Python Environment
Numpy	Version 1.21.5
Pandas	Version 1.4.4
Open-CV Python	4.6.0

Table 5.2 - Hardware Specifications for this project

5.3 HyperParameters

This section shows the hyperparameters that we have used in all of our models. This hyperparameters were manually selected after getting the best results out of them.

Parameter	Chosen Value
Loss Function	Categorical Cross Entropy
Optimizer	Adam
Epochs	100
Metric	Accuracy
Learning Rate	1e - 3
Batch Size	64

Table 5.3 - HyperParameters for MobileNet Transfer Learning Model

Parameter	Chosen Value
Loss Function	Categorical Cross Entropy
Optimizer	Adam
Epochs	100
Metric	Accuracy
Learning Rate	0.001
Batch Size	64

Table 5.4 - HyperParameters for CNN Model

Parameter	Chosen Value
Loss Function	Cross Entropy Loss
Optimizer	Adam
Epochs	500
Metric	Accuracy
Learning Rate	0.005
Batch Size	64

Table 5.5 - HyperParameters for Attentional Convolution Network

Parameter	Chosen Value
Epoch	200
Batch Size	250
Loss Function	Categorical Cross Entropy
Optimizer	Adam
Metric	Accuracy

Table 5.6 - HyperParameters for LSTM network for EEG based emotion detection

Chapter 6: Result & Analysis

This chapter presents the performance results of the model architectures used on the FER2013 dataset. In each case, the dataset has been split to training, validation and test set. The models have been trained on the training set, validated on the validation set, and finally the accuracy has been reported on the test set.

6.1 Evaluation Metrics used

To determine the best model, they were assessed against a variety of performance indicators. We evaluated the models' performance using three different metrics. They are: Accuracy, Precision, Recall.

Accuracy:

Accuracy defines the percentage of predictions the model has been able to properly predict against the total number of predictions (correct plus false).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.1}$$

where TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative

Precision:

Precision is the ratio between the number of correctly predicted positive observations to the number of total predicted positive observations. In other words, it attempts to measure what proportion of positive identifications was actually correct.

$$Precision = \frac{TP}{TP + FP} \tag{6.2}$$

Recall:

Recall is the ratio of correctly predicted positive observations to the all the predictions of all the class. In other words, it attempts to measure what proportion of actual positives was identified correctly.

$$Recall = \frac{TP}{TP + FN} \tag{6.3}$$

6.2 Evaluation Graphs used

We have also plotted the graphs of model Loss over epochs, Area Under the ROC Curve (AUC), Precision-Recall (PR) curve and confusion matrix to visualize our models' performance.

Loss Function Graph:

For all three of the models, we have used categorical cross entropy as the loss function. This is because this is a multi-class classification problem as we have different categories of emotions as output of the models. The equation of categorical cross entropy loss is as follows:

$$CE\ Loss = -\sum_{i=1}^{N} y_i \bullet \log(\hat{y}_i) \qquad (6.4)$$

Where N = size of the output, y_i = y_true (actual labels), $\hat{y}_i$ = y_pred (predicted labels)

Area Under the ROC Curve (AUC):

The Area Under the ROC Curve (AUC) measures the entire two-dimensional area underneath the entire ROC curve. The ROC curve, short for Receiver Operating Characteristic curve, is a graph which plots the performance of a model at all thresholds and plots TPR (True Positive Rate) on the y-axis vs. FPR (False Positive Rate) on the x-axis at different classification thresholds.

$$TPR = \frac{TP}{TP + FN} \qquad (6.5)$$

$$FPR = \frac{FP}{FP + TN} \qquad (6.6)$$

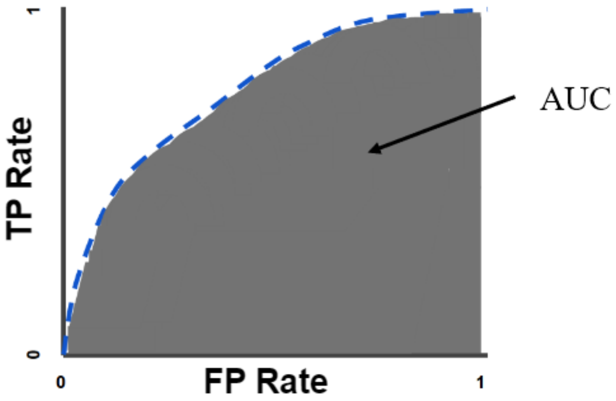


Figure 6.1 - AUC Curve

Precision-Recall (PR) curve:

A Precision-Recall (PR) curve is a plot of Precision on the y-axis vs Recall on the x-axis. A tradeoff between recall and precision can be seen through this graph at different threshold values. The greater the area under the PR curve, the higher is the precision and recall values, where high precision means a low false positive (FP) rate, and high recall means a low false negative (FN) rate. Therefore, the higher the area under the PR curve, the better is the model's performance.

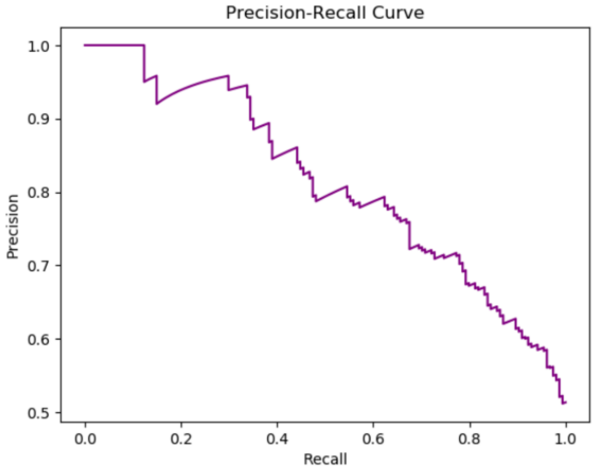


Figure 6.2 - Precision-Recall Curve

Confusion Matrix:

Confusion Matrix is a plot that is used to display the performance of a classifier model on the test data. The true values of the test data are known, and predicted values of the model are observed.

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

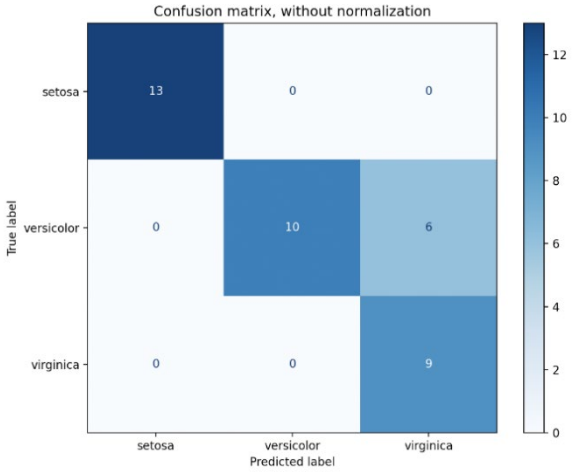


Figure 6.3 - A Confusion Matrix

6.3 Evaluation of MobileNet Transfer Learning Model

For the Transfer Learning method which uses MobileNet as the backbone, the entire 28,709 images of the FER2013 dataset in the training set were used to train the model, 3.5k images in the validation set were used to validate the model, and the model accuracy has been reported on the test data set. This model of ours achieved an accuracy of around 95% on the validation set. Figure 6.4 illustrates the confusion matrix of this model on predicted class vs actual class. Figure 6.5 shows all the graphs of this model's performance.

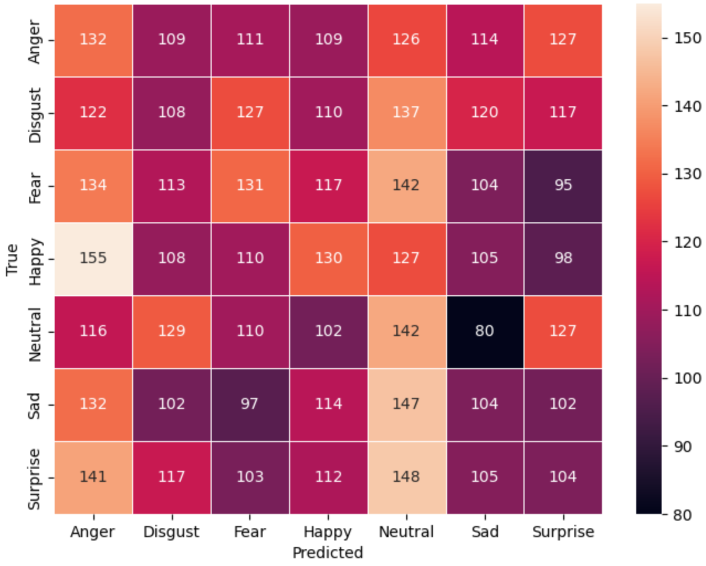


Figure 6.4 - Confusion Matrix of MobileNet Transfer Learning Model

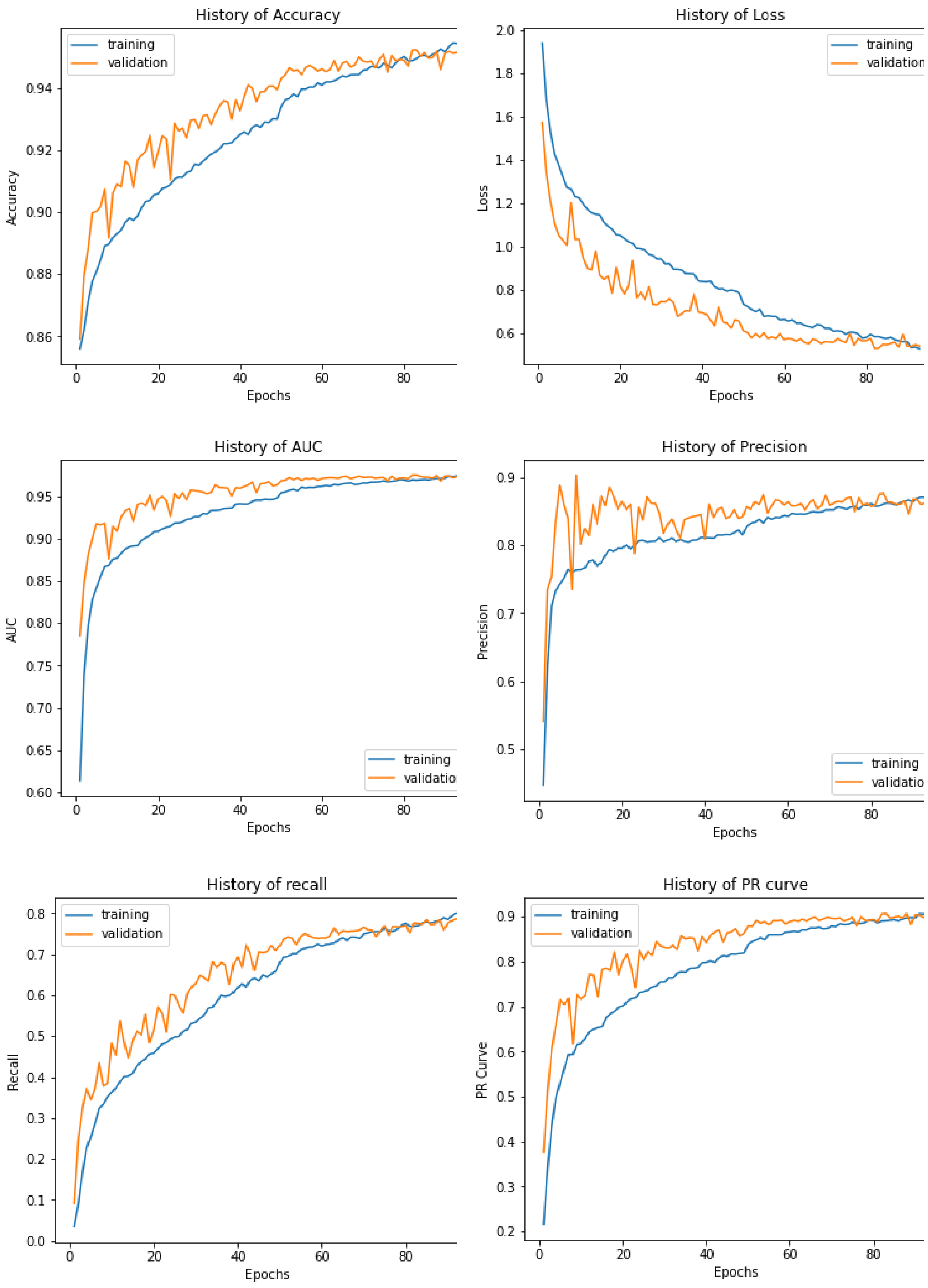


Figure 6.5 - Graphs of MobileNet Model

6.4 Evaluation of Convolutional Neural Network Model (CNN)

The CNN model has been also trained on the entire 28,709 images of the FER2013 dataset in the training set, 3.5k images in the validation set were used to validate the model, and the model accuracy has been reported on the 3,589 images in the test set. This model achieved an accuracy of 75.4% on the test set. Figure 6.6 illustrates the confusion matrix of this model. Figure 6.7 shows all the graphs of this model's performance.

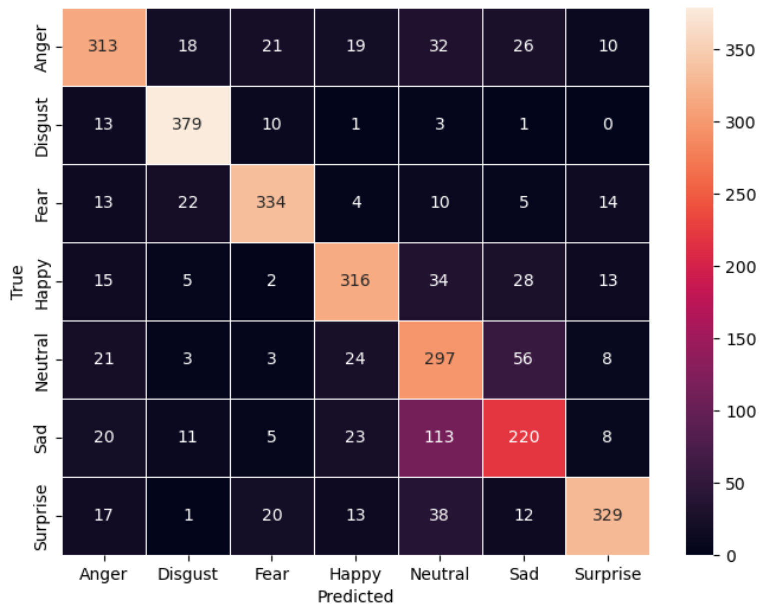


Figure 6.6 - Confusion Matrix of CNN Model

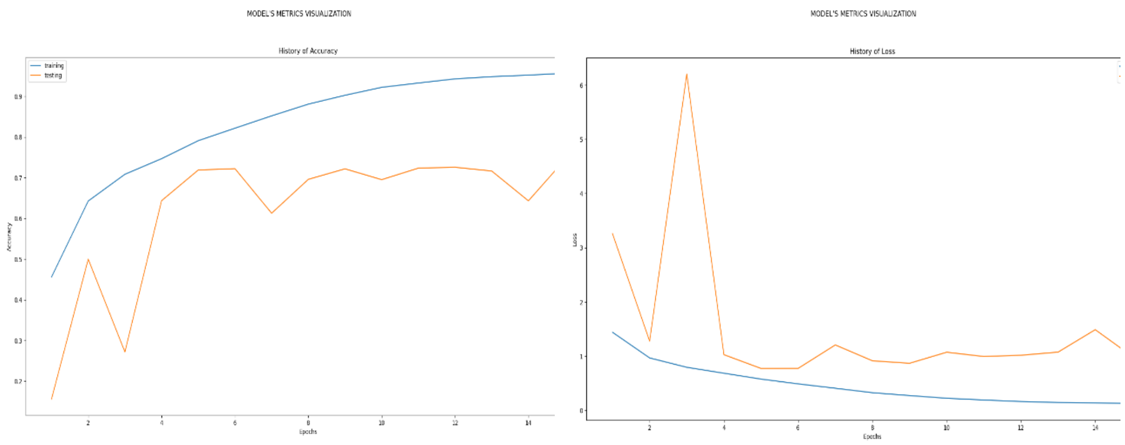
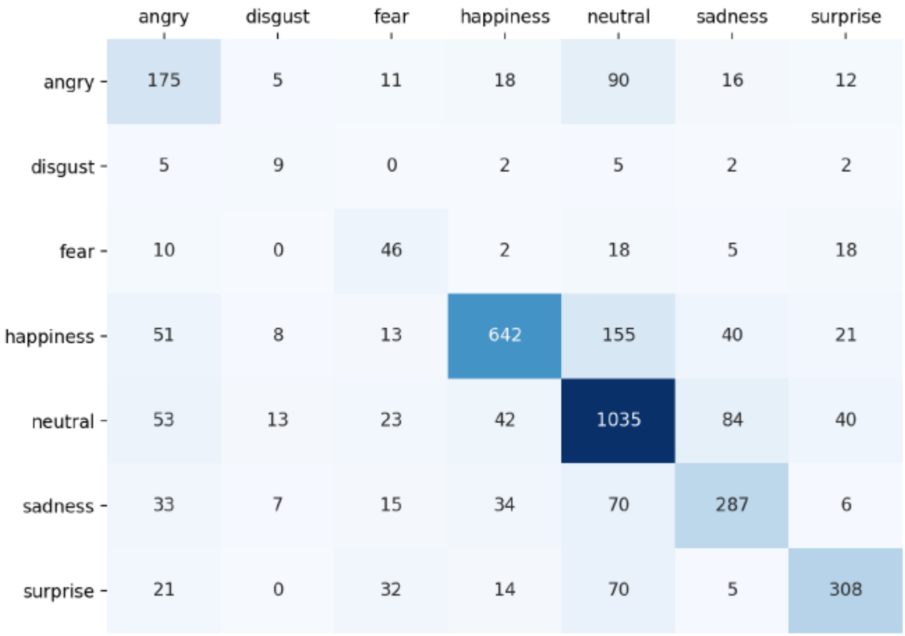


Figure 6.7 - Graphs of CNN Model

6.5 Evaluation of Attentional Convolution Network

This model has been also trained on the entire 28,709 images of the FER2013 dataset in the training set, 3.5k images in the validation set were used to validate the model, and the model accuracy has been reported on the 3,589 images in the test set. This attention based convolutional model achieved an accuracy of around 58%. Figure 6.8 shows the confusion matrix of this model.



	angry	disgust	fear	happiness	neutral	sadness	surprise
angry	175	5	11	18	90	16	12
disgust	5	9	0	2	5	2	2
fear	10	0	46	2	18	5	18
happiness	51	8	13	642	155	40	21
neutral	53	13	23	42	1035	84	40
sadness	33	7	15	34	70	287	6
surprise	21	0	32	14	70	5	308

Figure 6.8 - Confusion Matrix of Attentional CNN model

Chapter 7: Real-Time Live Emotion Detection

Real-time emotion detection using the trained model is no easy task as every frame from the input video needs to be sent to the model to predict, and the model then gives live feedback of the predicted emotion of the face, all happening simultaneously. For this to work properly and give out accurate emotion, we have used our CNN based trained model. A front-end interface was designed using open-cv-python library which takes live video and draws the bounding box and emotion on the face. Haarcascade Frontal Face Detector has been used to detect and track the face from the video input. The real-time live emotion detection design has been illustrated in Figure 7.1.

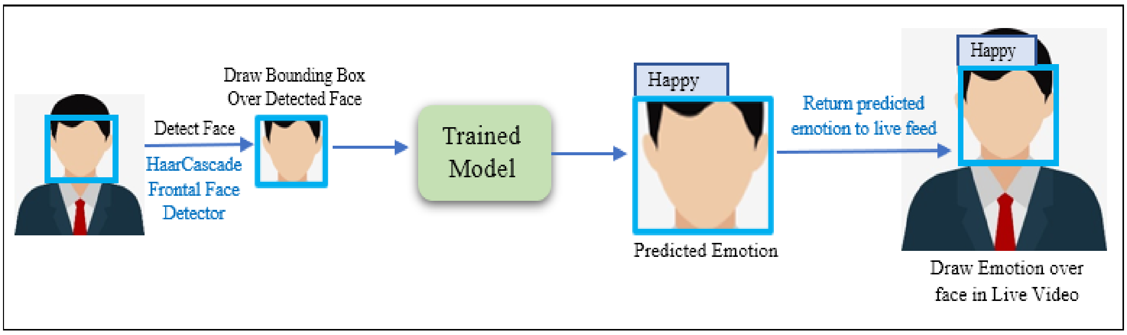

Draw Bounding Box Over Detected Face
Detect Face
HaarCascade Frontal Face Detector
Trained Model
Happy
Predicted Emotion
Return predicted emotion to live feed
Happy
Draw Emotion over face in Live Video

Figure 7.1 - Design of Real-Time Live Emotion Detection

Some predictions of real face expression videos downloaded from the internet using our trained model is shown in Figure 7.2. Real-Time Live emotion detection sample predictions of one of the authors face expressions is shown in Figure 7.3.

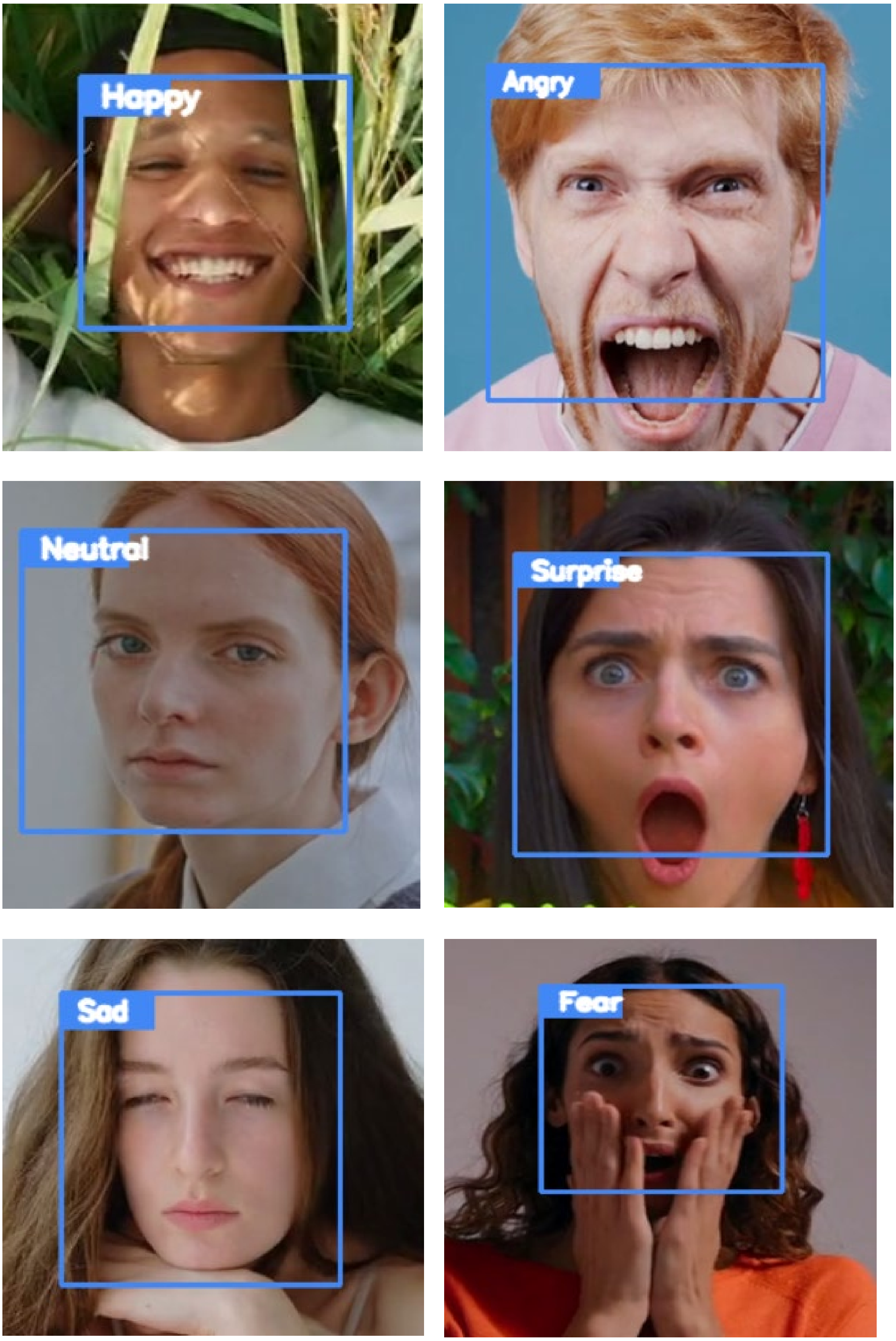


Figure 7.2 - Emotions predicted on different videos downloaded from the internet

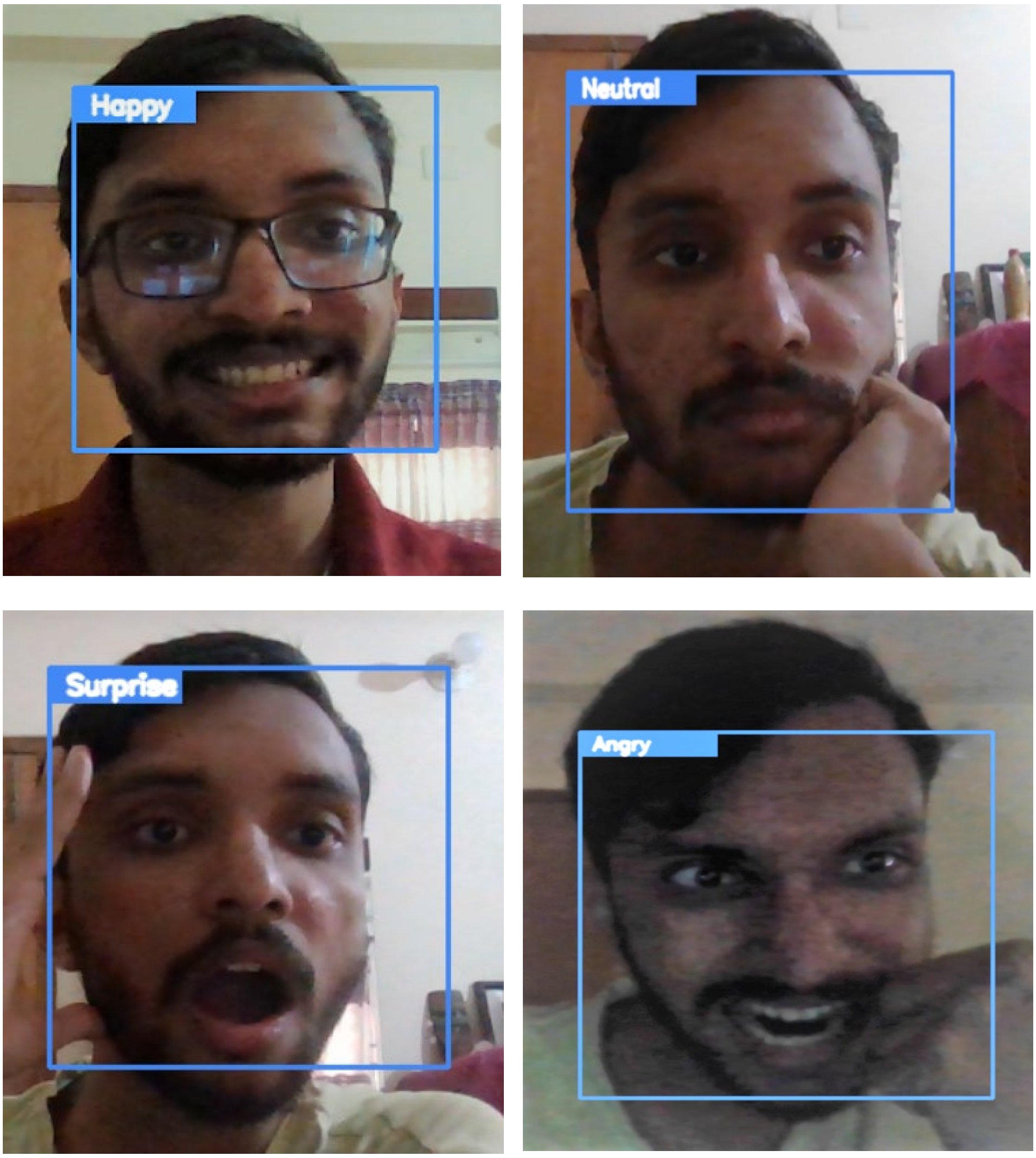


Figure 7.3 - Real-Time Emotion Detection Predictions

Chapter 8: Project Impact

8.1 Impact on the Research Community

A subfield of artificial intelligence and natural language processing that focuses on the identification, interpretation, and processing of human emotions is known as emotion analysis, sometimes known as affective computing or sentiment analysis. It has significantly impacted the scholarly community and received a lot of attention recently.

In psychology and mental health, emotional mapping has made a significant contribution. Researchers can gather information about how people feel and respond to various stimuli by looking at the emotional content of text or voice. This can help comprehend the emotional states of specific people or groups as well as for creating actions that will aid people in overcoming complicated feelings or enhancing their emotional well-being. For instance, researchers have studied the effects of social media on mental health, identified anxiety and depression predictions, and developed therapies to assist patients in coping with stress as well as unpleasant emotions using changes in a person.

A branch of artificial intelligence and natural language processing called emotion analysis, commonly referred to as affective computing or sentiment analysis, focuses on identifying, perceiving, and interpreting human emotions. It has recently drawn a lot of attention and had a big impact on the academic community.

The study of emotions has had a big impact on marketing and customer service in addition to these other fields. Businesses can better understand how their customers receive their products or services and make changes to increase customer satisfaction by evaluating customer feedback and reviews. The success of marketing initiatives can also be increased by using emotion analysis to determine which messages are most likely to connect with a given audience. For instance, researchers have utilized emotion analysis to examine how various forms of advertising affect consumer behavior, uncover factors that predict client loyalty, and create plans to increase customer happiness.

Emotion analysis has been applied in many other professions and fields, including psychology and marketing. For instance, it has been used in education and healthcare to examine the effects of various treatment modalities on student learning and engagement and in political science to examine the effects of political messages on public opinion.

Researchers in the field of emotion analysis are currently attempting to overcome several significant obstacles. Creating more precise and trustworthy ways of identifying and understanding emotions is one of the challenges. Even though much progress has been achieved in this area, there is always room for improvement, especially when recognizing softer or more complex emotions. Dealing with the subjectivity of emotions is another difficulty because different people may experience and express emotions in various ways.

This might make it challenging to correctly identify and categorize emotions, especially when there needs more context or other relevant information.

Despite these obstacles, emotion analysis has many possible uses and has already impacted the research community. The field will probably continue to influence several fields and disciplines as it develops significantly. The development of more sophisticated methods for analyzing emotional content in text and speech and the expansion of data availability are primarily to blame for this. Researchers will be able to design more potent interventions and methods for enhancing mental health, marketing, and a wide range of other areas as these approaches continue to advance since they can get deeper insights into the emotional states of individuals and groups.

Healthcare is one area where emotion analysis has affected the scientific community. By assisting medical personnel in comprehending their patients' emotional states, emotion analysis has the potential to enhance patient outcomes. For instance, healthcare workers can spot possible indications of sadness, anxiety, or other mental health issues and take the necessary action to address them by examining the emotional content of patient talks. This could entail counseling or directing patients to mental health specialists for additional treatment and evaluation.

Emotion analysis has been utilized in the corporate sphere to increase customer satisfaction and its uses in healthcare. Companies can better understand how consumers receive their goods or services by examining their comments and thoughts.

8.2 Social Impact

Emotion analysis may be advantageous for social media for several reasons:

- Emotion analysis can be used by organizations and businesses to determine how customers feel about a given sound, service, or occasion. This can help monitor client satisfaction and locate potential improvement areas.
- Addressing customer complaints and concerns: Businesses may quickly discover and respond to customer complaints or concerns by assessing the emotional content of social media postings and comments.
- Monitoring the emotional tone of online discussions: Using emotion analysis, one can keep tabs on the general emotional climate of online discussions on a given subject or problem. This can help discover potential trends or sentiment patterns and understand how users feel about a specific topic.
- Evaluation of marketing campaign effectiveness: Businesses can learn more about the kinds of communications most successful at evoking pleasant emotions and increasing engagement by examining how consumers emotionally react to marketing campaigns.

- Enhancing customer experience: Businesses can better satisfy the preferences and needs of their target market by customizing their products, services, and communication techniques following consumer emotions and needs. This can enhance overall client satisfaction and foster repeat business.
- The restrictions and possible ethical ramifications of using this technology must be considered. It is vital to keep in mind that emotion analysis is only sometimes accurate.

Chapter 9: Conclusion

Emotions serve an integral part of our lives. In this work, we have depicted human emotions from visual and EEG data using computer vision and deep learning technologies. We used two state-of-the-art datasets for our model training purposes – the DEAP dataset and the FER2013 dataset. The DEAP dataset consisted of emotion in the continuous valence-arousal-dominance scale, frontal face video and EEG signals of the participants. The FER2013 dataset consisted of 35,798 grayscale face expression images of 48x48 pixels, and each image falling inside a class of 7 emotions - Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral.

Using the DEAP dataset, emotion detection was done using EEG signals of the dataset. For this an LSTM architecture has been used to predict the emotions out of the EEG signals. The LSTM model gave an accuracy of 87%. On the other hand, using the FER2013 dataset, emotion detection was done on face expression images. For this, three architecture models were designed – a MobileNet Transfer Learning based model, CNN model, and an Attentional Convolution network. The MobileNet model achieved an accuracy of around 95%, the CNN model achieved an accuracy of 75.4% on the test data and the Attentional Convolution based network achieved an accuracy of around 58% on the test data.

Real-time emotion detection has been implemented finally. For this, a front-end interface was designed using OpenCV-python library. For emotion detection model to work, the face must first be detected accurately. Hence, Haarcascades Frontal Face detector was used to detect faces from the input frame. A bounding box is drawn around the detected face. The detected face is then passed onto the trained model which then predicts the emotion of that particular face. The emotion is then showed on top of the bounding box. All this happens simultaneously, as every frame from the real time live video gets passed onto the model and the predicted emotion is shown instantly.

References

[1] K. Cherry, "5 Reasons Emotions Are Important," Very Well Mind, 22 July 2022. [Online]. Available: https://www.verywellmind.com/the-purpose-of-emotions-2795181. [Accessed 10 December 2022].

[2] A. Banafa, "What is Affective Computing?," OpenMind BBVA, 06 June 2016. [Online]. Available: https://www.bbvaopenmind.com/en/technology/digital-world/what-is-affective-computing/#:~:text=Affective%20computing%20is%20the%20study,%2C%20psychology%2C%20and%20cognitive%20science. [Accessed 12 December 2022].

[3] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine,* vol. 18, no. 1, pp. 32 - 80, February 2001.

[4] T. B. Sheridan, "Human–Robot Interaction: Status and Challenges," *Human Factors: The Journal of the Human Factors and Ergonomics Society,* vol. 58, no. 4, pp. 525-532, 2016.

[5] D. Aneja, A. Colburn, G. Faigin, L. Shapiro and B. Mones, "Modeling Stylized Character Expressions via Deep Learning," *Asian Conference on Computer Vision,* vol. 10112, pp. 136-153, 2017.

[6] J. Edwards, H. J. Jackson and P. E. Pattison, "Emotion recognition via facial expression and affective prosody in schizophrenia: a methodological review," *Clinical psychology review,* vol. 22, no. 6, p. 789–832, 2002.

[7] H.-C. Chu, W. Tsai, M.-J. Liao and Y.-M. Chen, "Facial emotion recognition with transition detection for students with high-functioning autism in adaptive e-learning," *Soft Computing,* vol. 22, no. 5, pp. 1-27, 2017.

[8] C.-H. Chen, J. Lee and L.-Y. Lin, "Augmented reality-based self-facial modeling to promote the emotional expression and social skills of adolescents with autism spectrum disorders," *Research in Developmental Disabilities,* vol. 46, pp. 396-403, 2015.

[9] S. Hickson, N. Dufour, A. Sud, V. Kwatra and I. Essa, "Eyemotion: classifying facial expressions in VR using eyetracking," *IEEE Winter Conference on Applications of Computer Vision (WACV),* pp. 1626-1635, 2019.

[10] M. A. Assari and M. Rahmati, "Driver drowsiness detection using face expression recognition," in *Proceedings of the 2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, Kuala Lumpur, Malaysia, 2011.

[11] C. Tao and K.-H. Yap, "Discriminative BoW Framework for Mobile Landmark Recognition," *IEEE Transactions on Cybernetics,* vol. 44, no. 5, pp. 695-706, 2013.

[12] A. Mollahosseini, D. Chan and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016.

[13] P. Liu, S. Han, Z. Meng and Y. Tong., "Facial expression recognition via a boosted deep belief network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[14] B. G¨okberk, A. A. Salah and L. Akarun, "Rank-based decision fusion for 3D shape-based face recognition," in *Proceedings of the International Conference on Audio-and Video-Based Biometric Person Authentication*, Hilton Rye Town, NY, USA, 2005.

[15] K. Han, D. Yu and I. Tashev., "Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine," in *Fifteenth annual conference of the international speech communication association*, Singapore, Malaysia, 2014.

[16] C.-H. Wu, Z.-J. Chuang and Y.-C. Lin., "Emotion recognition from text using semantic labels and separable mixture models," *ACM Transactions on Asian Language Information Processing ,* vol. 5, no. 2, pp. 165-183, 2006.

[17] V. Pavlovic, R. Sharma and T. S. Huang, "Visual Interpretation of Hand Gestures for Human-Computer Interaction A Review," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 677-695, 1997.

[18] P. Petrantonakis and L. Hadjileontiadis, "Emotion Recognition From EEG Using Higher Order Crossings," *IEEE transactions on information technology in biomedicine,* vol. 14, no. 2, pp. 186-197, 2010.

[19] A. Mehrabian, "Communication without words," in *Communication Theory*, London, UK, Routledge, 2017, pp. 193-200.

[20] P. Ekman, "An Argument for Basic Emotions," *Cognition and Emotion,* vol. 6, no. 3, pp. 169-200, 1992.

[21] C. Calistra, "The Universally Recognized Facial Expressions of Emotion," Kairos, 15 March 2015. [Online]. Available: https://www.kairos.com/blog/the-

universally-recognized-facial-expressions-of-emotion. [Accessed 20 December 2022].

[22] S. Koelstra, C. Muehl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt and I. Patras, "DEAP: A Database for Emotion Analysis using Physiological Signals," *EEE Transactions on Affective Computing,* vol. 3, no. 1, pp. 18-31, 2012.

[23] S. A. a. K. R. a. o. Inigo, "Real-Time Emotion Recognition System using Facial Expressions and Soft Computing methodologies," *Journal of Science Technology and Research (JSTAR),* vol. 3, no. 1, 2022.

[24] A. I. a. S. N. F. a. A. A. D. a. E.-S. A. a. F. E. a. S. A. Siam, "Deploying Machine Learning Techniques for Human Emotion Detection," *Computational Intelligence and Neuroscience,* vol. 2022, 2022.

[25] A. T. Kaviya P, "Group Facial Emotion Analysis System Using," in *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)*, 2020.

[26] I. a. S. A. R. a. B. M. E. Lasri, "Facial Emotion Recognition of Students using Convolutional Neural Network," in *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*, 2019.

[27] G. K. a. T. U. S. Verma, "Affect representation and recognition in 3D continuous valence--arousal--dominance space," *Multimedia Tools and Applications,* vol. 76, no. 2, pp. 2159-2183, 2017.

[28] G. Y. G. C. H. L. a. W. L. XIANZHANG PAN, "A Deep Spatial and Temporal Aggregation framework for Video-based Facial Expression Regression," *IEEE Access,* vol. 7, pp. 48807-48815, 2019.

[29] W.-L. a. L. B.-L. Zheng, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Transactions on autonomous mental development,* vol. 7, no. 3, pp. 162-175, 2015.

[30] L. Trainor, "Frontal brain electrical activity (EEG) distinguishes valence and intensity of musical emotions," *Cognition & Emotion - COGNITION EMOTION,* vol. 15, pp. 487-500, 2001.

[31] M. Z. H. &. S. R. Yuhao Zhang, "DeepVANet: A Deep End-to-End Network for Multi-modal Emotion Recognition," in *Human-Computer Interaction -- INTERACT 2021*, 2021.

[32] H. AG, Z. M, C. B, K. D, W. W, W. T, A. M and A. H., " Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint ,* vol. abs/1704.04861, 2017.

[33] A. Sarkar, "Understanding depthwise separable convolutions and the efficiency of MobileNets," Medium, Jun 2012. [Online]. Available: https://towardsdatascience.com/understanding-depthwise-separable-convolutions-and-the-efficiency-of-mobilenets-6de3d6b62503#:~:text=MobileNet%20is%20a%20CNN%20architecture,on%20mobile%20and%20embedded%20devices..

[34] B. Zhu, P. Hofstee, J. Lee and Z. Al-Ars, "An Attention Module for Convolutional Neural Networks," *Artificial Neural Networks and Machine Learning – ICANN 2021.,* vol. 12891, p. 167–178, 2021.

[35] J. Brownlee, "A Gentle Introduction to the Rectified Linear Unit (ReLU)," Machine Learning Mastery, 20 August 2020. [Online]. Available: https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/. [Accessed 22 December 2022].

[36] J. Brownlee, "A Gentle Introduction to Pooling Layers for Convolutional Neural Networks," Machine Learning Mastery, 5 July 2019. [Online]. Available: https://machinelearningmastery.com/pooling-layers-for-convolutional-neural-networks/#:~:text=Maximum%20pooling%2C%20or%20max%20pooling,the%20case%20of%20average%20pooling.. [Accessed 22 December 2022].