Face Expression and Movement Based Pattern Recognition

Pooja Yadav   
MTech-Artificial IntelligenceDelhi Technological University *Delhi, India*  
poojayadav\_24afi21@dtu.ac.com

*Abstract*—The classification of facial expressions and movements is a significant area of study in the fields of computer vision and human-computer interaction. This project explores the use of convolutional neural networks (CNNs) to accurately recognize and classify facial expressions from a diverse dataset of different facial movement and expressions. We utilize the MobileNetV2 architecture, known for its efficiency and accuracy in image classification tasks, as the base model for feature extraction. The dataset comprises images categorized into seven distinct facial expressions: surprise, sad, neutral, happy, fear, disgust, and angry.

The methodology involves preprocessing the images to a standardized size of 224x224 pixels and augmenting the dataset to enhance model robustness. The MobileNetV2 model is fine-tuned by adding fully connected layers followed by a Softmax layer to predict the probability distribution over the seven classes. ReLU activation functions are employed to introduce non-linearity, and the model is trained using the Adam optimizer with a sparse categorical cross-entropy loss function.

Experimental results demonstrate the effectiveness of the proposed approach in accurately classifying facial expressions, with notable improvements in recognition accuracy compared to traditional methods. The findings highlight the potential applications of this technology in areas such as emotion detection, security systems, and human-computer interaction. Future work includes exploring the integration of temporal dynamics to enhance movement classification and improve real-time performance.

Keywords—MobileNetV2, Activation Function, Adam Optimizer, cross-entropy loss, haarCascade\_frontal\_default.

# INTRODUCTION

[Facial expression recognition](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/facial-recognition) (FER) is to separate the specific expression state from the given static image or video to determine the psychological emotions of the recognized object, the realization of the computer's understanding and recognition of facial expressions have fundamentally changed the relationship between human and computer, to achieve better [human computer interaction](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/human-computer-interaction) (HCI).

Facial expression and movement classification has become an essential aspect of computer vision and human-computer interaction. The ability to accurately recognize and classify facial expressions can significantly enhance applications in fields such as emotion detection, security systems, and interactive user interfaces.

This project aims to implement a deep learning-based approach to classify seven distinct facial expressions: surprise, sad, neutral, happy, fear, disgust, and angry.

To achieve this, we utilize the MobileNetV2 architecture, known for its efficiency and accuracy in image classification tasks, as the backbone of our model. The dataset comprises images from various categories corresponding to different facial expressions. Each image undergoes preprocessing, including resizing and colour conversion, to standardize the input for the neural network.

The model architecture includes convolutional layers for feature extraction, followed by fully connected layers and a Softmax layer for classification. The convolutional layers leverage ReLU activation functions to introduce non-linearity, and optional pooling layers are used to down sample the feature maps. The fully connected layers further refine the extracted features, and the softmax layer outputs the probability distribution over the seven classes.

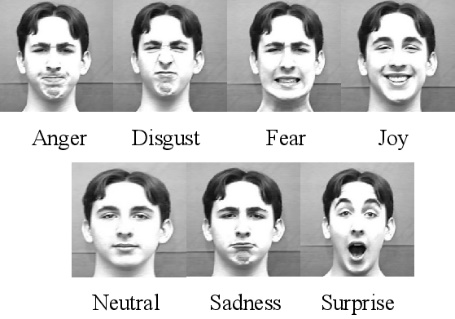
According to this paper's analysis of facial expression traits, there aren't many distinctions between people's expressions when they are the same, meaning that there aren't many differences across classes. When two people display the same expression, there are significant disparities between them; in other words, there are significant intra-class differences. Extracting fine-grained characteristics of local expression changes—like those in the mouth and eyes—is the key to FER. To give local expression features more attention, the lightweight MobileNetV1 model was modified to include an attention module. In the meantime, model parameters for expression recognition are optimized using the central loss function and softmax loss. The Softmax function decreases the cross-problem between classes while increasing their distance from one another. Internal variations within the same emoticon are minimized via the central loss function.

II. DATASET USED (FER-2013)

## The goal of this project is to develop a deep learning model that can accurately classify facial expressions into one of seven categories:

## Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral.

## We will use convolutional neural networks (CNNs) to train our model, as they have been shown to be effective in image classification tasks. Our approach involves training a CNN model from scratch on the FER-2013 dataset.



*Fig:1 Different Expressions*

The FER-2013 dataset consists of 48x48 pixel grayscale images of faces that have been automatically registered to be centred and occupy a similar amount of space in each image. The dataset contains 24,400 images, with 22,968 examples in the training set and 1,432 examples in the public test set.

* Images: The dataset contains approximately 30,000 facial RGB images of different expressions.
* Image Size: All images are resized to 48x48 pixels.
* Expressions: The images are labelled with one of seven facial expressions: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral.
* Label Distribution: The distribution of images across different expressions is relatively balanced, with each expression having around 5,000 samples, except for Disgust, which has about 600 samples.

## Preprocessing of Dataset

To train a deep learning model for human emotion recognition using the FER-2013 dataset, we will use both data generators and data augmentation in this section. It would be impractical to put every image into memory at once due to the dataset's size and complexity. As a result, we will be able to effectively train our model on the complete dataset by using data generators to create batches of photos live during training.   
  
Additionally, we will be using a variety of data augmentation methods, including flipping, shifting, and rotating, on the training images. Our deep learning model's performance may be enhanced by expanding the size and variety of our training set. By adding fluctuation to Overfitting, which happens when the model gets too tightly fitted to the initial training set and performs badly on fresh, unknown data, can be avoided throughout the training process with the aid of data augment.

*Fig: Facial Expressions of FER-2013 Dataset*

*Fig:3 Dataset Classification*

## Classification of Facial Expressions

Facial expressions provide information

about emotions, convey our intentions and goals, and play an

essential role in human interaction. The ability to recognize

and understand facial expressions automatically facilitates the

intended communication.

The process in the classiﬁcation of human facial expressions

consists of three stages: face detection, feature extraction,

and facial expression classiﬁcation. In this study, the authors

applied a system that could classify facial expressions at a

macro level, consisting of 7 (seven) basic human expression

Facial expressions provide information

about emotions, convey our intentions and goals, and play

essential role in human interaction. The ability to

recognize and understand facial expressions facilitates

automatically the intended communication.

The process in the classiﬁcation of human facial

expressions consists of three stages: face detection,

face extraction and facial expression classiﬁcation.

In this study, the authors applied a system that could

classify facial expressions at a macro level, consisting of

7(seven) basic human expressions:

### Happy:

A smile expression is an expression that can show that someone is feeling happiness or liking something. The happy expression is on the upward movement of the cheek muscles and the sides or edges of the lips to form a smile.

*Fig:4 Facial Expression happiness*

*2) Anger*:

Anger facial expressions arise from the match between what is expected and a reality. The expression is shown on both sides of the inner eyebrows which are merging and leaning down, while the lips are narrowing, and the way the eyes are sharp when looking.

*Fig:5 Facial Expression anger*

*3) Sadness:*

A face that shows sadness appears when disappointment or a feeling of missing something. Based on the characteristics of a sad facial expression when the eye loses focus, the lips are pulled downwards, and the upper eyelid droops.

*Fig:6 Facial Expression Sadness*

*4) Fear:*

The form of expression that appears when someone experiences an inability to cope with any event or in a scary atmosphere, then that person is said to be afraid. The expression of fear on a person’s face is seen from the two eyebrows that rise at the same time, the eyelids tighten, and the lips that are open horizontally.

*Fig:7 Facial Expression fear*

*5) Disgust:*

A person who expresses his face in a state of disgust due to seeing something not common or listening to information that is not worth hearing. An expression of disgust will be read when a person’s face in the area of the nose bridge is wrinkled and the upper lip rises.

*Fig:8 Facial Expression disgust*

*6) Surprise:*

Expressions of surprise are obtained when someone does not know beforehand an event or message received that is sudden, unexpected or important. Expression is a shocked face represented by the raised eyebrows, the eyes wide open, and the mouth opening reflex.

*Fig:9 Facial Expression surprise*

*7) Neutral:*

The facial expression of a person who is identified as being neutral and has no assumption of what that person is about to do or doing. The expression is shown when there is no sense of movement in the lip or face points.

*Fig:10 Facial Expression Neutral*

## HaarCascade\_frontalface\_default Data

The **HaarCascade\_frontalface\_default** is a pre-trained classifier in OpenCV for detecting frontal faces. It uses machine learning with a cascade function trained from numerous positive and negative images. This XML file identifies faces in images by applying the classifier to detect regions that match the trained features, drawing rectangles around detected faces.

The Haar-Cascade Face Detection Algorithm is a sliding-window type of algorithm that detects objects based upon its features.

* *Haar Face Features:* The Haar-Cascade model, employs different types of feature recognition that include the likes of
* Size and location of certain facial features. To be specific, nose bridge, mouth line and eyes.
* Eye region being darker than upper-cheek region.
* Nose bridge regio being brighter than eye region.

Intel's 'haarcascade\_frontalface\_default.xml'

This 'XML' [file](https://github.com/opencv/opencv/blob/master/data/haarcascades/haarcascade_frontalface_default.xml) contains a pre-trained model that was created through extensive training and uploaded by Rainer Lienhart on behalf of Intel in 2000.

# METHODOLOGY

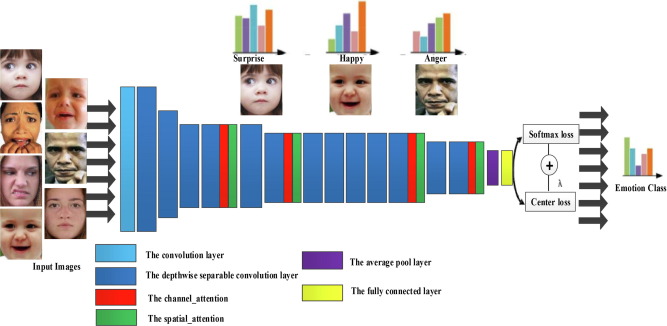
## CNN Architecture:

Convolutional Neural Network (CNN) is an algorithm included in the deep neural network or Deep Learning family due to the high network depth and is significantly superior when implemented on image data. Deep Learning (DL) is a Neural Networks technique that uses specific techniques such as Restricted Boltzmann Machine (RBM) to accelerate the learning process in Neural Network, which uses multiple layers or more than seven layers. With the existence of Deep Learning, the time required for training will be less because the problem of losing the gradient in backpropagation will be lower.

The initial research underlying the CNN discovery was first carried out by Hubel and Wiesel, who carried out visual cortex studies of the cat visual sense. The visual cortex in animals is very powerful in existing visual processing systems until many studies have been inspired by how it works and has produced new models such as Neocognitron, HMAX LeNet-5, and AlexNet.

The CNN method results from the development of the Multilayer Perceptron (MLP) method for two-dimensional processing data, for example, images or sounds. The way CNN works is similar to MLP, but on CNN, each neuron present in two dimensions. Unlike MLP, where each neuron is only one dimension. Image processing can be started as a specific feature, such as brightness or edge increasing complexity on features that uniquely define objects according to layer thickness. In general, the existing layer types in Convolutional Neural Network divide into two, namely Extraction Layers Features (Feature Extraction Layer) and Classification Layer (Classification Layer).

The CNN layers have a 3-dimensional arrangement of neurons (width, height, depth). Width and height are the sizes. Convolutional Neural Network Algorithm of the layers, while depth refers to the number of layers. A CNN can have tens to hundreds of millions of layers, each of which learns to detect various images. Image processing applies to each training image at a different resolution, and the output of each image is processed



*Fig:11 CNN Architecture*

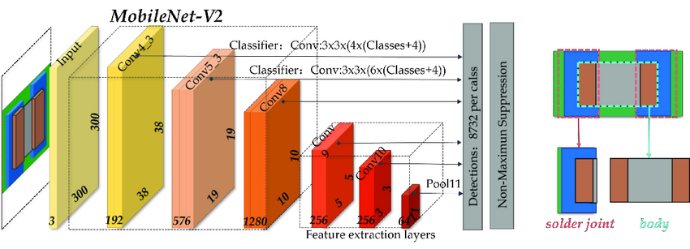
## Introduction to MobileNetV2:

Following the success of AlexNet, which utilized a convolutional neural network (CNN) for image classification and secured first place in the ImageNet competition in 2012, researchers have developed an array of deep neural network models, including the well-known VGGNet16/19, GoogleNet, and ResNet50. These models have demonstrated superior performance compared to traditional classification algorithms. However, as the complexity of these networks has increased, the substantial storage requirements and computational demands associated with model calculations have begun to restrict the applicability of deep learning models. Conventional CNNs require significant memory and computational resources, rendering them unsuitable for mobile and embedded devices.

In response to this challenge, Google introduced MobileNetV1, a lightweight deep neural network designed specifically for mobile applications. This CNN features a reduced model size, fewer trainable parameters, and lower computational demands, thereby optimizing resource utilization while maximizing model accuracy.

The core idea of MobileNetV1 network is to replace the standard [convolution operation](https://www.sciencedirect.com/topics/computer-science/convolution-operation) with depthwise separable convolution (DSC) to reduce model parameters. Specifically, DSC is to use the 3 × 3 [convolution kernel](https://www.sciencedirect.com/topics/computer-science/convolution-kernel) with only one layer thickness, sliding layer by layer on the input tensor, and generate an output channel after each convolution. When the convolution was completed, use 1 × 1 pointwise convolution to adjust the thickness, as

shown in Fig.



*Fig:12 MobileNet V2 Architecture*

However, the MobileNetV1 introduces a ReLU [activation function](https://www.sciencedirect.com/topics/computer-science/activation-function) after the deep [convolution layer](https://www.sciencedirect.com/topics/computer-science/convolution-layer), which cannot change the number of channels. The extracted features are single channel, and the operation of the ReLU in the convolution layer output with fewer channels will lead to information loss.

*Fig: 13 MobilenetV2*

## Libraries:

TensorFlow:

* TensorFlow is an end-to-end open-source platform for machine learning. It is used here for creating and training the neural network model.
* *Keras*, which is a high-level API of TensorFlow, is used for building and compiling the model.

OpenCV:

* OpenCV (Open Source Computer Vision Library) is an open-source computer vision and machine learning software library. It is used here for image processing tasks, such as reading images, resizing images, and converting color spaces.

Matplotlib:

* Matplotlib is a plotting library for the Python programming language. It is used here for displaying images.

NumPy:

* NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. It is used here for handling arrays and performing numerical operations.

PyTorch:

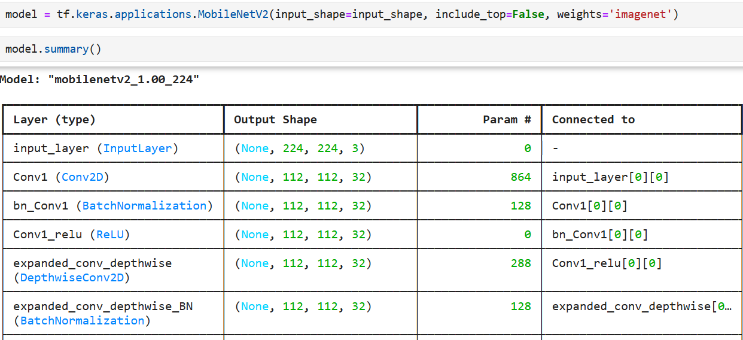
* PyTorch is an open-source deep learning framework developed by Facebook's AI Research lab (FAIR). It's widely used for developing machine learning and deep learning applications

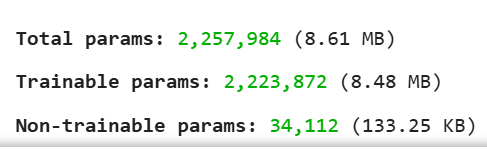
# RESULT AND ANALYSIS

The FER-2013 dataset is structured as follows:

1. **Training Set**: Contains **28,709 images**.
2. **Test Set**: Contains **7178 images**.

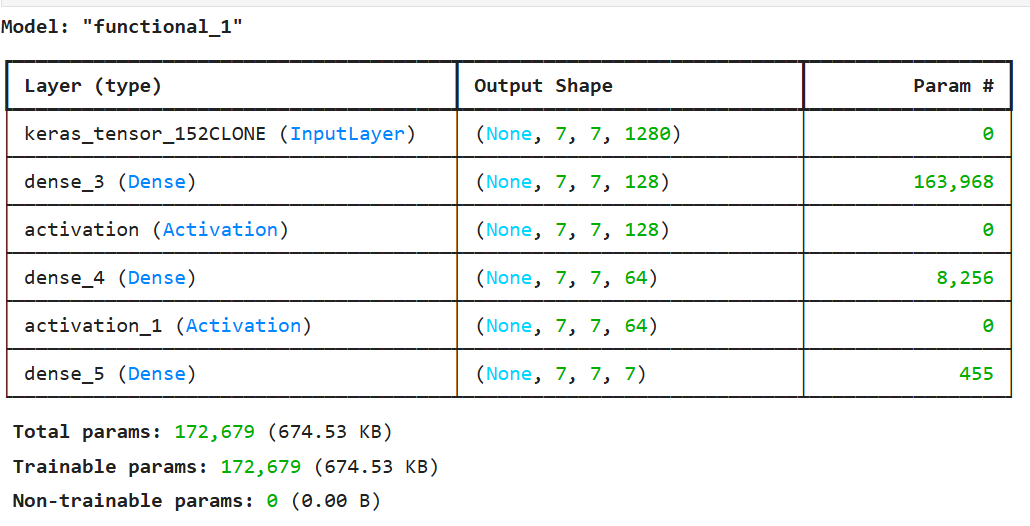
## MobileNetV2 Model with parameters:

Total Parameters:



*Fig 14& 15: Parameters in MobilenetV2*

## Model Parameters with layers:

**

*Fig: 16 Model Described*

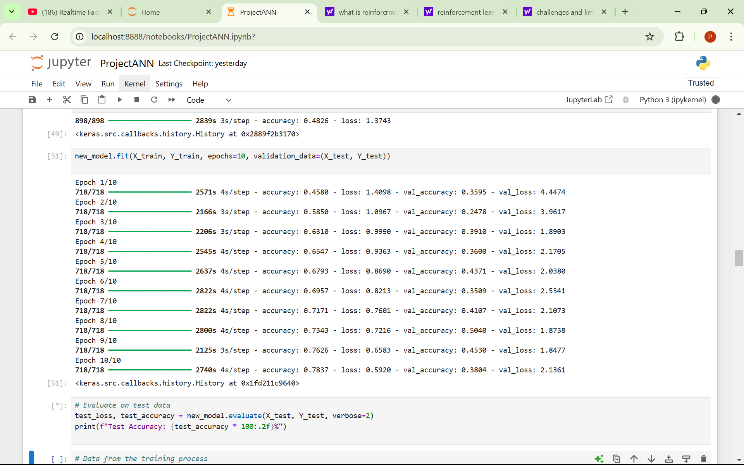
## Accuracy and Loss over Epochs:

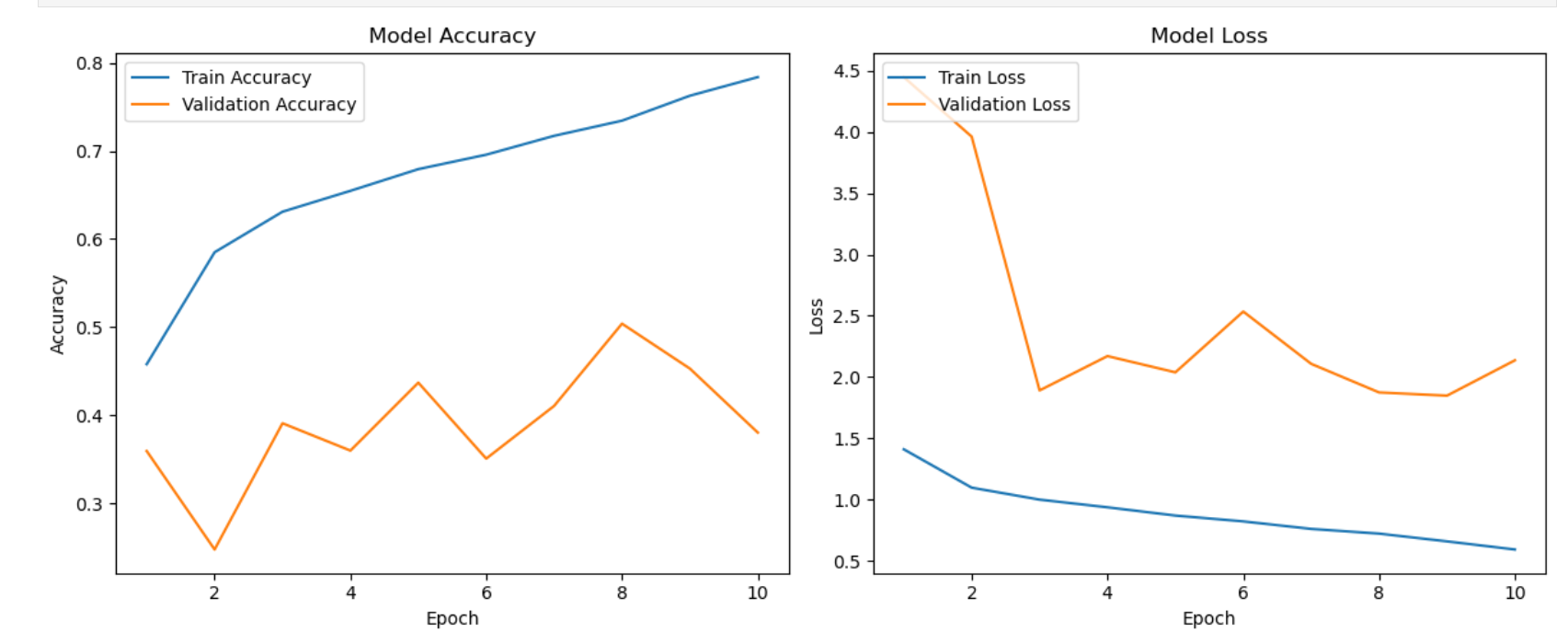
As the number of epochs increases, the accuracy get increases for facial expression detection while loss keeps on decreasing for test data set after training the model on training dataset of all expressions.

epochs = [1, 2, 3]

accuracy = [0.4824, 0.5930, 0.6237]

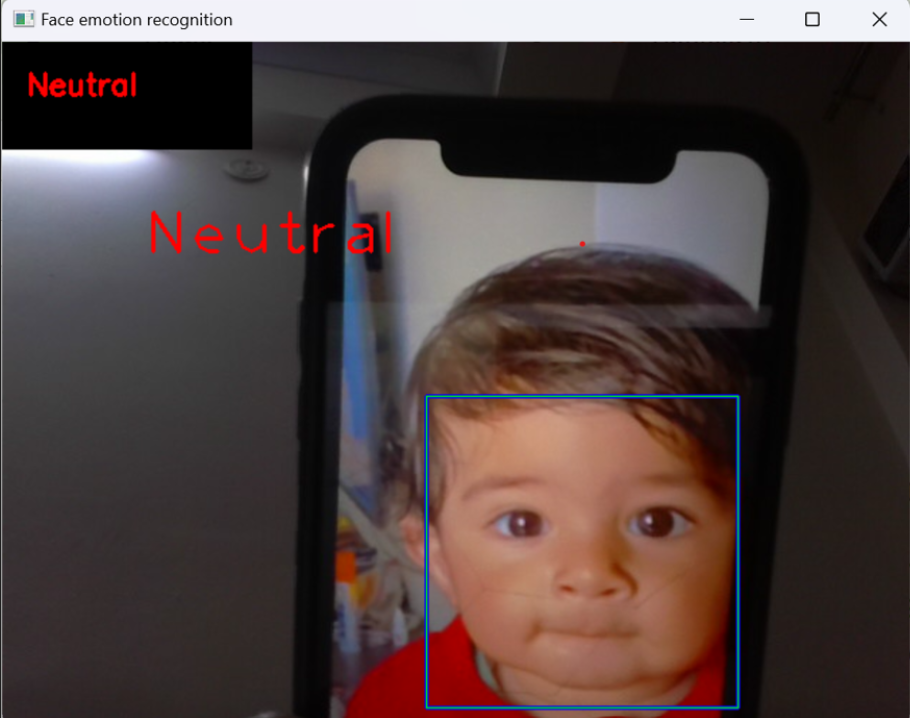
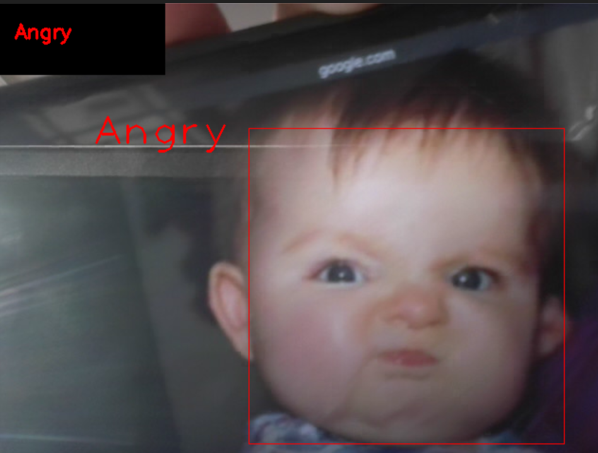
loss = [1.3640, 1.0847, 1.0075]





*Fig: 17 Accuracy &Loss V/s Epochs*

## Face Expressions:



*Fig: 18 Different Facial Expressions Snippet*

## Predictions of Class:

## 

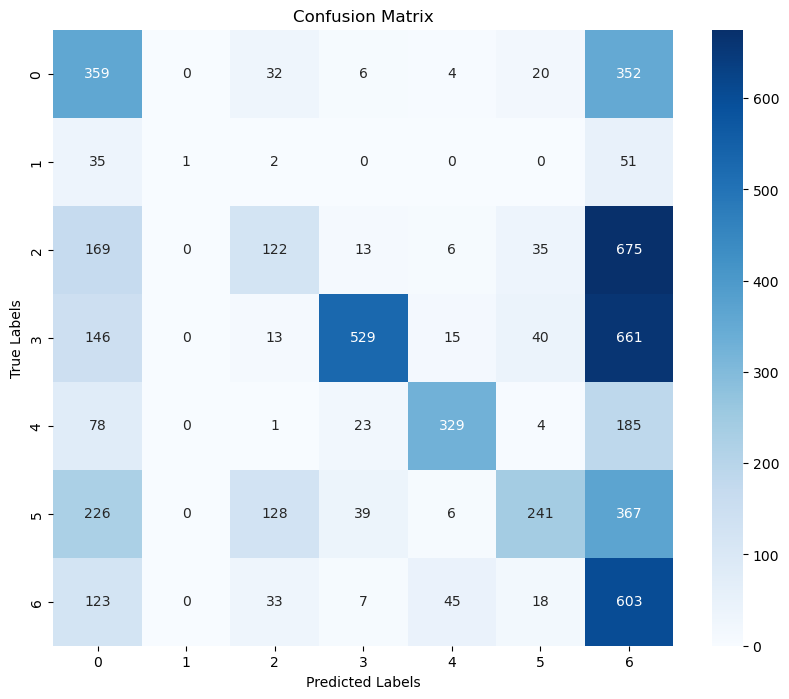
## 

*Fig 19. Random data testing*

* Prediction=0 class (Happy)
* Prediction=4 class (Surprise)

## Evaluation Matrix(Confusion matrix):

Confusion matrix visualizes the performance of a classification model:



*Fig 20. Confusion matrix*

Insights:

* The model performs well for certain classes (like 3 and 6) but struggles with distinguishing others (like 2 and 6, or 0 and 6).
* This imbalance may indicate a need for further model tuning, better features, or more data for the confused classes.

# Future Work And Scope

Improvement of Accuracy

* Advanced Architectures: Explore more sophisticated neural network architectures such as ResNet, EfficientNet, or attention-based models to improve classification accuracy.
* Ensemble Methods: Implement ensemble methods combining predictions from multiple models to enhance robustness and accuracy.

Data Augmentation and Expansion

* Augmented Data: Utilize advanced data augmentation techniques to create a more diverse training dataset, improving model generalization.
* Larger Datasets: Incorporate additional datasets from sources like AffectNet, CK+, or custom datasets to enrich the training data and improve performance on diverse expressions.

Real-Time Application

* Optimization: Optimize the model for deployment on edge devices like smartphones or embedded systems for real-time facial expression detection.
* Latency Reduction: Implement techniques to reduce inference time and ensure seamless real-time performance.

# CONCLUSION

The outcomes of the conducted trials indicate that a system has been successfully designed, providing a comprehensive overview of the research subject through the application of the Convolutional Neural Network (CNN) methodology. This system is capable of predicting seven distinct human facial expressions by employing Facial Emotion Recognition (FER) techniques, utilizing the FER-2013 dataset. The design of the system in this study encompasses the following processes:

1) The training phase employs the Facial Expression Recognition 2013 Dataset (FER-2013) and implements the Convolutional Neural Network (CNN) approach for feature extraction and effective facial expression prediction.

2) Real-time facial expression recognition is achieved by detecting facial objects through the Haar Cascade method, with the Convolutional Neural Network (CNN) utilized for the classification of facial expressions.

3) During the facial expression recognition process, the detected expressions are displayed in the information viewer on the expression display board.

Future work will be based on following factors to have enhanced working model with more features:

1) Implement a more complex architecture by incorporating additional convolutional layers and fully connected layers with optimal configurations, resulting in a deeper CNN structure that enhances accuracy.

2) Consider utilizing a new dataset or increasing the volume of training data to achieve improved accuracy during testing, particularly for scenarios involving distances greater than 5 meters, varying viewing angles, and rotated images.

3) Select hardware, such as a high-resolution digital camera equipped with autofocus capabilities, to ensure clear imaging even when the subject is in motion, thereby enhancing network performance prior to model integration for effective detection and recognition.

# ACKNOWLEDGEMENT

I extend my deepest gratitude to my mentors for their guidance, and to my family and friends for their unwavering support. Delhi technological university fully supported this project research collaborated with the Department of Computer Science Engineering in Artificial intelligence. Special thanks to my project coordinator for their collaboration, and to the resources from online communities that enriched this work. This project wouldn't have been possible without everyone's contributions and encouragement

# REFERENCES

1. Dynamic face authentication systems: Deep learning verification for camera close-Up and head rotation paradigms, https://www.sciencedirect.com/science/article/abs/pii/S0167404822000281
2. <https://medium.com/@pragyasingh18rathore/face-expressions-and-movement-based-pattern-matching-user-authentication-system-f56d7d1de9fc>
3. A-MobileNet: An approach of facial expression recognition- https://www.sciencedirect.com/science/article/pii/S1110016821006682
4. https://www.researchgate.net/publication/378885866\_Effective\_Facial\_Expression\_Recognition\_System\_Using\_Machine\_Learning
5. The Facial Emotion Recognition (FER-2013) Dataset for Prediction System of Micro-Expressions Face Using the Convolutional Neural Network (CNN) Algorithm based Raspberry Pi- https://www.researchgate.net/publication/347959965\_The\_Facial\_Emotion\_Recognition\_FER-2013\_Dataset\_for\_Prediction\_System\_of\_Micro-Expressions\_Face\_Using\_the\_Convolutional\_Neural\_Network\_CNN\_Algorithm\_based\_Raspberry\_Pi
6. https://viswalahiri.github.io/Face\_Detection\_Haar\_Cascade/scripts/
7. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
8. <https://www.geeksforgeeks.org/what-is-mobilenet-v2/>
9. https://www.cs.cmu.edu/~rsalakhu/papers/oneshot1.pdf
10. <https://github.com/opencv/opencv/blob/4.x/data/haarcascades/haarcascade_frontalface_default.xml>

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**