Emotion Recognition from Facial Images using Deep Learning Architectures

Author

Arfa Fatima Yaseen

319333

Supervisor

Dr. Arslan Shaukat

DEPARTMENT OF COMPUTER ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

October 2021

Emotion Recognition from Facial Images using Deep Learning Architectures

Author

Arfa Fatima Yaseen

319333

A thesis submitted in partial fulfillment of the requirements for the degree of

MS Computer Engineering

Thesis Supervisor

Dr. Arslan Shaukat

Thesis Supervisor's Signature: ____

DEPARTMENT OF COMPUTER ENGINEERING
COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY, ISLAMABAD
October, 2021

Declaration

I certify that this research work titled “*Emotion Recognition from Facial Images using Deep Learning Architectures”* is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources it has been properly acknowledged / referred.

Signature of Student

Arfa Fatima Yaseen

319333

Language Correctness Certificate

This thesis has been read by an English expert and is free of typing, syntax, semantic, grammatical and spelling mistakes. Thesis is also according to the format given by the university.

Signature of Student
Arfa Fatima Yaseen
319333

Signature of Supervisor
Dr.Arslan Shaukat

Copyright Statement

- Copyright in text of this thesis rests with the student author. Copies (by any process) either in full, or of extracts, may be made only in accordance with instructions given by the author and lodged in the Library of NUST College of E&ME. Details may be obtained by the Librarian. This page must form part of any such copies made. Further copies (by any process) may not be made without the permission (in writing) of the author.
- The ownership of any intellectual property rights which may be described in this thesis is vested in NUST College of E&ME, subject to any prior agreement to the contrary, and may not be made available for use by third parties without the written permission of the College of E&ME, which will prescribe the terms and conditions of any such agreement.
- Further information on the conditions under which disclosures and exploitation may take place is available from the Library of NUST College of E&ME, Rawalpindi.

Acknowledgements

All praise and glory to Almighty Allah (the most glorified, the highest) who gave me the courage, patience, knowledge and ability to carry out this work and to persevere and complete it satisfactorily. Undoubtedly, HE eased my way and without HIS blessings I can achieve nothing.

I would like to express my sincere gratitude to my advisor Dr. Arslan Shaukat for boosting my morale and for his continual assistance, motivation, dedication and invaluable guidance in my quest for knowledge. I am blessed to have such a co-operative advisor and kind mentor for my research.

Along with my advisor, I would like to acknowledge my entire thesis committee Dr. Muhammad Usman Akram and Dr. Sajid Gul Khawaja for their cooperation and prudent suggestions.

My acknowledgement would be incomplete without thanking the biggest source of my strength, my family. I am profusely thankful to my beloved parents who raised me when I was not capable of walking and continued to support me throughout in every department of my life and my loving sisters who were with me through my thick and thin.

Finally, I would like to express my gratitude to all my friends and the individuals who have encouraged and supported me through this entire period.

Dedicated to my exceptional parents: ***Prof. Muhammad Yaseen & Farhat Iqbal*** *and adored sisters and friends whose tremendous support and cooperation led me to this accomplishment*

Abstract

Facial expressions (FE) or human countenance reflect psychological reactions or intentions stimulated within the mind in response to any social or personal event. These expressions play a significant role in conveying messages to the observer in non-verbal stealth mode. With the advancements in technology, facial expression recognition (FER) is considered crucial in understanding human behavior. We can infer those feelings and expressions are the essences of any interaction. In the same way, we need to make human machine interaction as communal as human-human interaction by making machines proficient at detecting human emotions by reading facial expressions. From recent year's discoveries, a Set of multiple features have been recognized that provide possibly useful outcomes in the field of emotion recognition. Few preprocessing steps have been performed on the image data set before the extraction of features. In this work different prevalent methodologies, techniques and the types of features that are used by researchers in the past to predict the facial expression over time will be combined, so that a new and more efficient model can be designed. The purpose of this research is to design an automated system which can recognize seven basic emotions of human namely anger, disgust, fear, happy, sadness, Neutral and surprise for effective communication between humans and computers. The single algorithm to provide perfect recognition in all the scenarios has never been established so far; however, the research has been in progress to develop substitutes or new models to improve the recognition process. A deep learning algorithm is explained in this research work for classifying the facial expression of the human. The proffered method investigates the effectiveness of deep convolution neural network (DCNN) with the help of multiple models, and the best achieved result is 94.88% of FER2013.

Key Words: *Facial expression recognition, efficientnetB0, deep convolutional neural networks, deep learning, VGG16, FER2013.*

Table of Contents

List of Figures

List of Tables

Chapter#1

Introduction

CHAPTER 1: INTRODUCTION

With the penetration of digital technology in our lives, the need for effective human-computer interactions has become an inevitable necessity. Most of the things we deal with in our routine activities involve smart digital systems that operate by observing verbal expressions, movement of body parts, and subtle facial gestures. With the progressive involvement of these systems in different verticals of life like security and surveillance using visual data, password unlocking in cell phones, video advertisements, games, etc. [1][2], the need to quantitatively describe the human behavior for digital manipulation has become the most attracted research topic over the past few years. Until now, the human expressions that researchers could have converted into digital parameters for machine learning are divided into seven categories that depict the emotions of Angriness, Repugnance, Fearfulness, Happiness, Neutral Reactions, Sadness and Surprise [3].

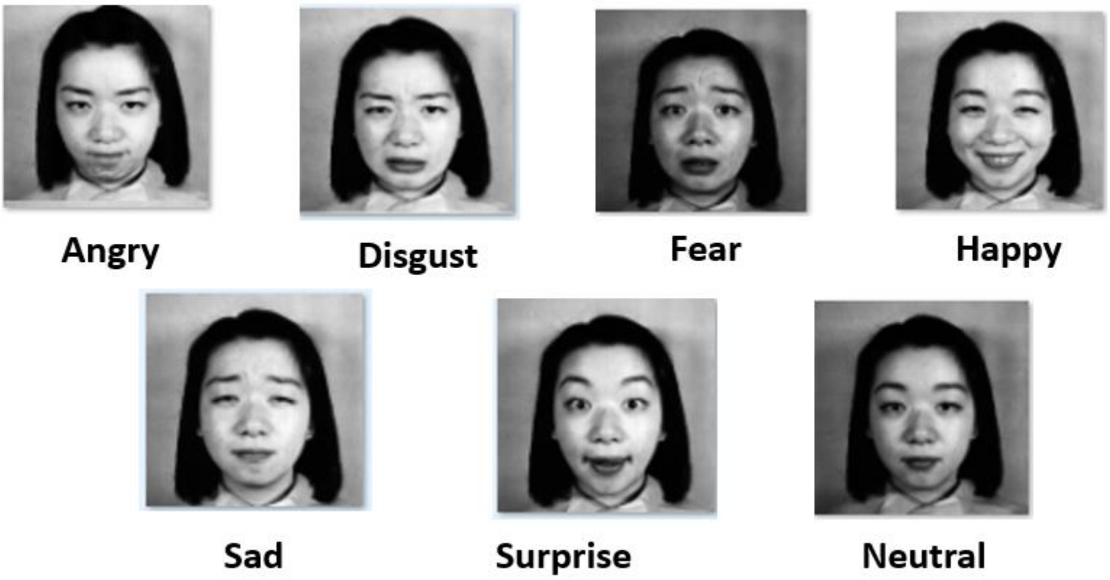


Figure 1.1: Seven Basic Facial Expression Sample

Among all the parts of the human body, face is not only considered as an identification source of a person but a mirror to his inner emotions and feelings. Where the future world is envisioned to be operated by digitally designed robots, the requirement to develop facial expression sensing tools has increased. Although humans can easily interpret each other emotions, making the machines behave like them is still a challenging task. Over the last few years many Facial Expression Recognition (FER) algorithms have been developed to serve the purpose. FER Systems usually involves stages like Face Detection, Feature Extraction and Emotion Classification [4].

With the development of an idea to design FER systems on the working principle of human brain, deep learning algorithms have come into scope and found to achieve great success as well as better accuracy over the traditional methods of expression detection. Deep learning strategy for expression detection also comprises different variants of algorithms with slightly different working mechanisms depending upon the nature of applications. One such variant acknowledged as Convolutional Neural Network (CNN) has gained popularity in the computer vision field due to its accuracy in producing required results for designing a robust CNN, the model is needed to be trained with numerous images representing the complete detail of the environment where the system is to be used.

Figure 1.2: Facial Expression Analysis [5]

Various feature extraction filters are then used to capture generic features and then operations like pooling and discarding are used to produce the desired results [5]. Moreover, the existing Graphics Processing Units (GPUs) speed up the training mechanism of deep neural networks to report processing time problems in testing phase and training phases. Despite many issues of recognition accuracy like lower resolutions, occlusion, variations in lighting conditions, and head pose variations have led to development of different state of art models, the limitations in producing the accurate results in real-time uncontrolled environment still need more work in this field.

1.1 Motivation

Many researchers are working to improve the recognition accuracy of facial expression recognition systems by handling some challenges [6], [7], [8]. The factors such as rapid head movement, lightening conditions, fading and shading from nearby objects, large number of similar objects or a number of objects belonging to the same category sometimes makes the detection more challenging and in these scenarios the datasets that have been designed under controlled environments seem to fail in training the system for real-time detection, thus causing the need to develop different data sets for each application. So far, one data set that can train the systems for all type of real-time applications hasn't been produced which is the major drawback in determining the accuracy of any FER detection model. The system that appears to work well with one dataset not necessarily produces the same results for the other application. The major problem that FER systems are facing is an insufficiency of training data in terms of quality and quantity. To overcome the above challenges, robust and reliable feature extraction techniques are required.

The work introduced in this thesis highlights the techniques to improve the detection accuracy of the system using deep learning. This research work deals with investigating of methods using deep learning techniques to deal with the issue of recognition accuracy of lower resolution images for expression recognition. The key factor of this research is to improve the recognition accuracy of the real-time facial expression dataset which contains real-world images with challenges and the laboratory trained dataset images that are trained in a controlled environment for the cross-database evaluation study.

The feature extraction process becomes more difficult in real-world images than the images trained in a controlled environment. The work done under this research is mainly influenced by the methods of feature extraction and classification used in [10], [11], [12], [13] and many more.

1.2 Problem Statement

The ability to predict the expression of face can surely help in several important applications such as the applications which require emotional intelligence; it can also be used in the domains like health care, mobile computing and many more. Figure 1.3 define how facial expression help in playing song on the basis of your mood. The work introduced in this research focuses on recognizing facial expressions from the images using deep learning techniques to improve its recognition accuracy. This research work deals with investigating of methods using deep learning techniques to deal with the issue of recognition accuracy of lower resolution images for facial expression recognition.

Figure 1.3: Facial Expression application in social media [9]

The key factor of this research is to improve the recognition accuracy of the real-time facial expression dataset which contains real-world images with challenges and the laboratory trained dataset images that are trained in a controlled environment for the cross-database evaluation study. The feature extraction process becomes more difficult in real-world images than the images trained in a controlled environment.

1.3 Aims and Objectives

Major objectives of the research are as follow:

1. Data Enhancement using technique of contrasting has been introduced to produce more discrimination between objects. The proposed model maps the intensity value of input image and produces improved output data.
2. To study and investigate existing feature-based techniques used in convolutional neural networks for improving recognition accuracy of a facial expression recognition system.
3. To test and train the proposed model, a new method has been introduced that makes use of less computations to produce the desired results thus improving time of training.
4. In order to test the performance of the model in practical scenarios both the competition benchmark and the datasets collected from the actual scenes are used to evaluate the results.

1.4 Structure of Thesis

This work is structured as follows:

Chapter 2 provides the overview of the research work accomplished in the field of FER so far and emotion recognition techniques and neural network.

Chapter 3 elaborates the details of proposed model and the training process. Section IV lists a series of conducted experiments and the comparison results with other state-of-the-art works.

Chapter 4 explain proposed methodology in detail

Chapter 5 lists a series of conducted experiments and the comparison results

Chapter 6 concludes the thesis and explain future work in facial expression field

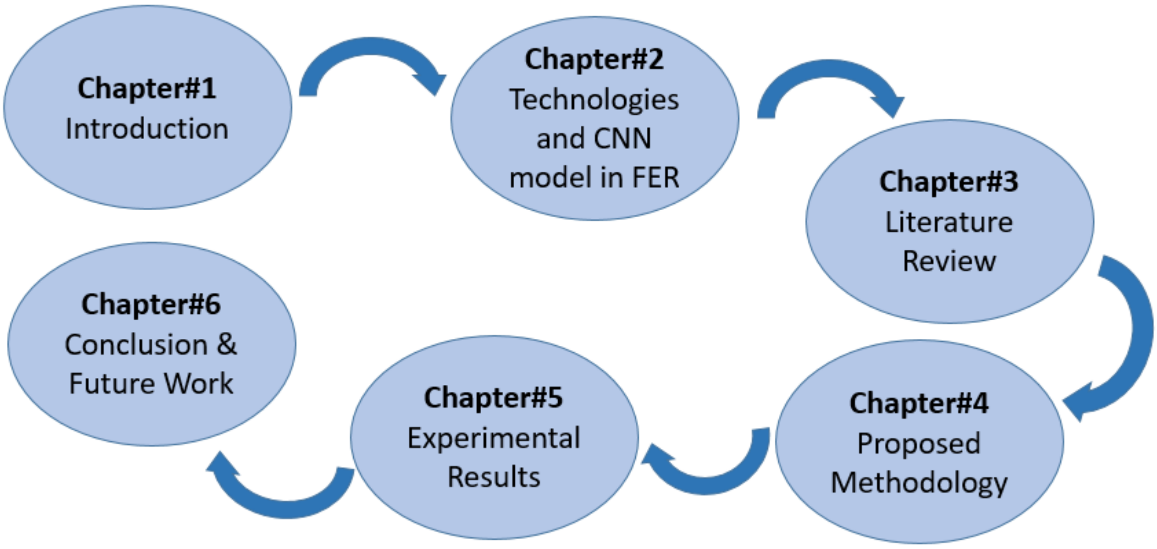


Figure 1.4: Thesis Layout

Chapter#2

Related Background

CHAPTER 2: RELATED BACKGROUND

Facial Expression recognition algorithms generally involve three steps i.e., (i) Face Detection, (ii) Feature Extraction, and (iii) Emotion Classification. In the first step, the features of face like eyes, nose, lips and mouth are detected. In the second step the interpretation of expressions is done using the shape and attributes of features extracted in the first step, and in the last step classifiers are used to label those expressions. For classification of expressions, training data is used to train the classifiers, and the efficiency of such classifiers to adapt to the real-time environment can be optimized using different techniques [14].

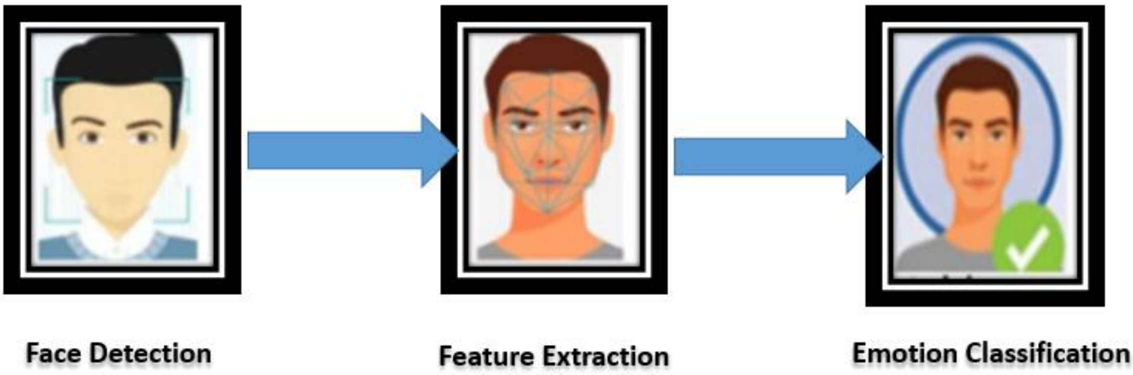


Figure 2.1: Facial Expression recognition algorithms

Facial expressions are usually subtle and require more deep interpretation rather classifying them into the basic emotional states of happiness, anger, surprise, fear etc. that requires the development of systems to analyze both facial expressions based on both permanent facial features (brows, eyes, mouth) and transient facial features (deepening of facial furrows).

2.1 Development of Facial Action Coding System (FACS)

In [15], Ekman introduced the technique of Facial Action Coding System (FACS) based on anatomy to interpret facial expressions from the movement of muscles of face. This technique involves classification of facial movements into different action units (AUs) (facial feature movement or combination of features movements) which are then used to categorize the facial expressions. They defined 44 AUs from which 30 AUs are based on analyzing the muscle movements, 12 for upper face, and 18 are for lower face. As an expression can be a combination of multi-feature movements, accordingly AUs can occur either singly or in combination. Initially these systems were found to be little effective in expression classification [16] due to certain issues like manual highlighting and defining of feature points on face area, restricted face movement as frequent head movement or face view from side causes inaccurate results, limited usefulness for large databases and real-time applications due to large number of calculations involved to accommodate combination of 44 AUs, difficulty in expression recognition when more subtle emotions appear on human face, etc.; thus, making them less efficient for real-time applications. Over the time, many automatic detection techniques have been developed to overcome the existing issues in handling the AUs.

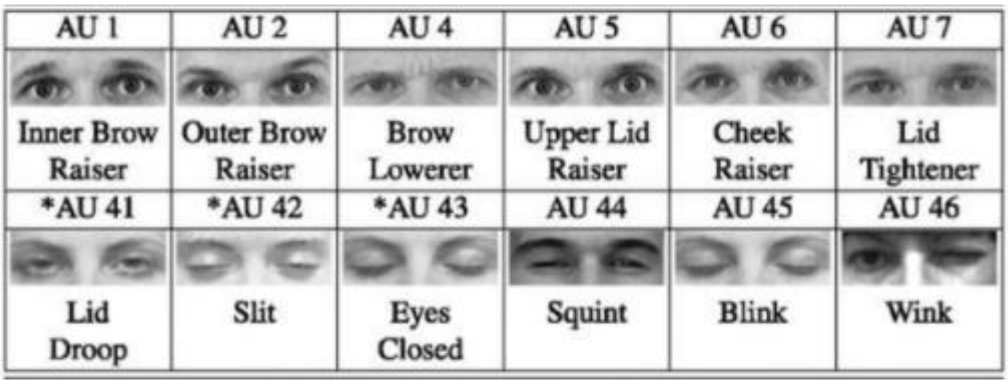


Figure 2.2: Upper face Action Units [17]

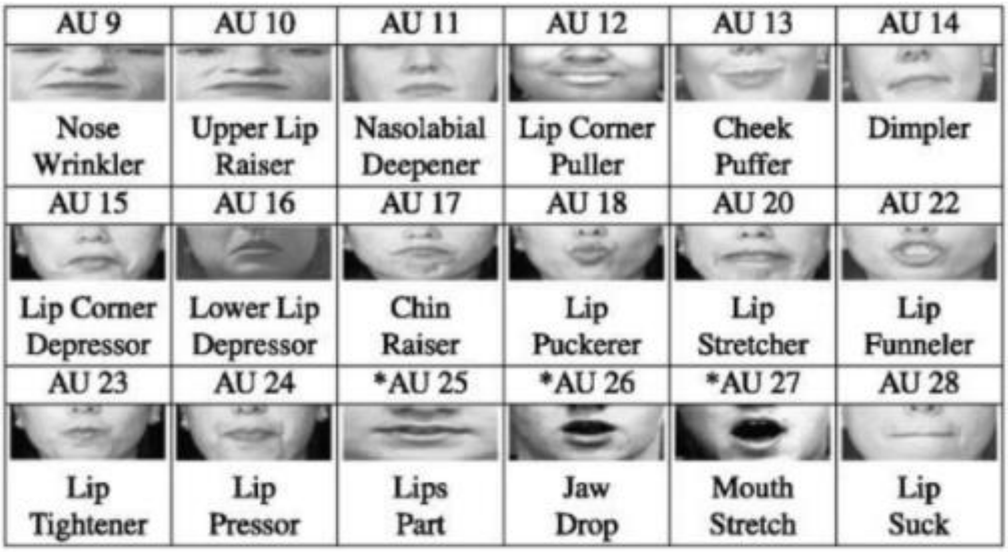


Figure 2.3: Lower face Action Units [17]

2.1.1 Incorporation of Automatic Expression Detection in FACS

The limitations of FACS techniques in time management, detection, and processing have enticed the researchers for the development of automatic detection systems for AUs. The automatic detection of AUs involves the incorporation of training data where the classification systems are first trained for dynamic changes detection before they are used to handle the real time changes [18].

2.2 Face Detection in FER

Detecting facial and non-facial areas in any image is the first step in image processing [19]. The algorithms normally begin with the detection of eyes following the detection of brows, nose, mouth, etc. The methods used in face detection can be knowledge-based, feature-based, template matching, or appearance based.

Knowledge-based methods involves the development of rules to describe a face; while feature based methods use basic features such as a eyes or nose to detect a face. Template-matching methods correlate the standard face patterns and extracted features to locate a face in the image. On the other hand, appearance-based methods make use of machine learning to find the relevant characteristics of face images.

Different classifiers based on these techniques to detect a face are discussed in subsequent sections

2.2.1 Viola-Jones framework

The variation in factors such as pose, expression, position and orientation, skin color and pixel values, presence of glasses and differences in camera gain, lighting conditions and image resolution often make face detection difficult. The concept of deep learning has helped a lot in eliminating such issues.

Computer vision researchers, Paul Viola and Michael Jones, in 2001 introduced a framework to detect faces in real-time applications. They made use of training data to help the model understand facial and non-facial objects [20]. The model used to store some features at each stage and correlate them with new features at the subsequent stages to detect a face. However, the framework has limitations for the scenarios in which faces are not clearly visible or covered.

2.2.2 Haar Cascade Classifier

Haar Classifier works on principal introduced by Voila and Jones to detect face in image or video. In this classifier, features can be detected by varying the size of the pixel group. This model offers the functionality to train itself during the training process by identifying a set of features which are most contributing for the face detection due to which it offers less computation complexity in subsequent stages [19], [21].

2.2.3 Adaptive Skin Color

Adaptive skin-color model locates skin color objects in the images to detect a face followed by a mechanism to differentiate between facial and non-facial objects. However, its performance gets affected with different levels of illumination due to which adaptive gamma corrective method is used to eradicate the issue. However, the computational complexity of corrective method, makes it less feasible to be used in real-time applications [19].

2.2.4 Adaboost Contour Points

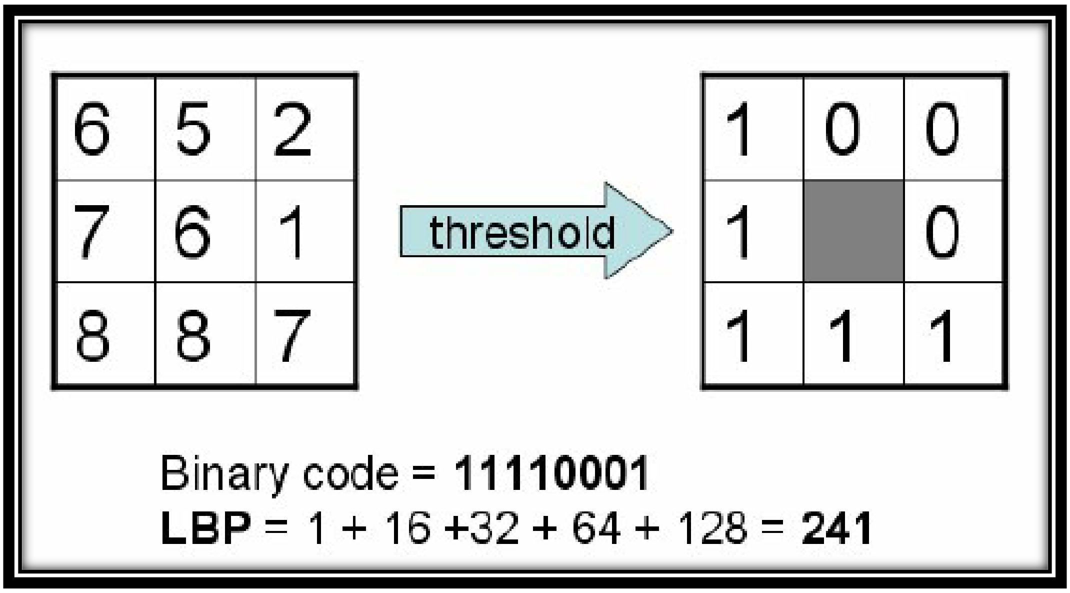
Adaboost makes use of a number of classifiers cascaded with each other resulting in a highly efficient classifier which is then used to detect a face. It offers low computational complexity and therefore suitable to detect face in a real-time environment. The low computational complexity of this classifier is also due to its property to detect contour points instead of all basic features which in turn offers less effort in obtaining necessary information from image or video [22].

2.3 Feature Extraction in FER

Feature extraction is the conversion of pixel data into shape, color and texture to represent a real object. Feature extraction usually acquires a large collection of data and keeps on filtering it to maintain a record of useful information only. It then categorizes the results into different emotion classes. The techniques used for feature extraction in FER are discussed in subsequent sections:

2.3.1 Local Binary Pattern (LBP)

The LBP calculates the brightness relationship between each pixel contained in the image and its local neighbourhood. LBP is used in computer vision for texture analysis [23]. Initially it was able to work on 3X3 matrices only that involved converting of image into grey scale and thresholding a center pixel value to its neighbors. The eight values around the central pixel value are called its neighbors. In this technique, a binary matrix is developed from the input matrix by comparing the central matrix value to each of its neighbors. If value (intensity) of central pixel is less than neighbor, the value is set to 0; and if it is greater than the neighbor, the value is set to 1. The results of this binary matrix are then stored in 8-bit array starting from either clockwise or anti clockwise around the 3x3 binary matrix. Figure 2.4 explain LBP method.



6	5	2
7	6	1
8	8	7

1	0	0
1		0
1	1	1

Binary code = **11110001**

LBP = 1 + 16 +32 + 64 + 128 = **241**

Figure 2.4: Local Binary Pattern (LBP) Algorithm

The array values are then converted into a single decimal value. The process is repeated for each possible 3x3 matrix in the input image and the results are stored in output decimal matrix. At the end, histograms are established from the decimal values to represent fine grained details extracted from image [6]. Now a day’s circular LBPs of varying radii are used to remove the limitations of square LBPs. This is formulated as below:

$$\boldsymbol{LBP\ p,r} = \sum_{p=0}^{p-1} \boldsymbol{S(g_p - g_c)\ 2^p}, \qquad \boldsymbol{S(x)} = \begin{matrix} \boldsymbol{1, x \geq 0} \\ \boldsymbol{0, x < 0} \end{matrix} \qquad \ldots\ldots\ldots\ldots\ldots (\boldsymbol{2.1})$$

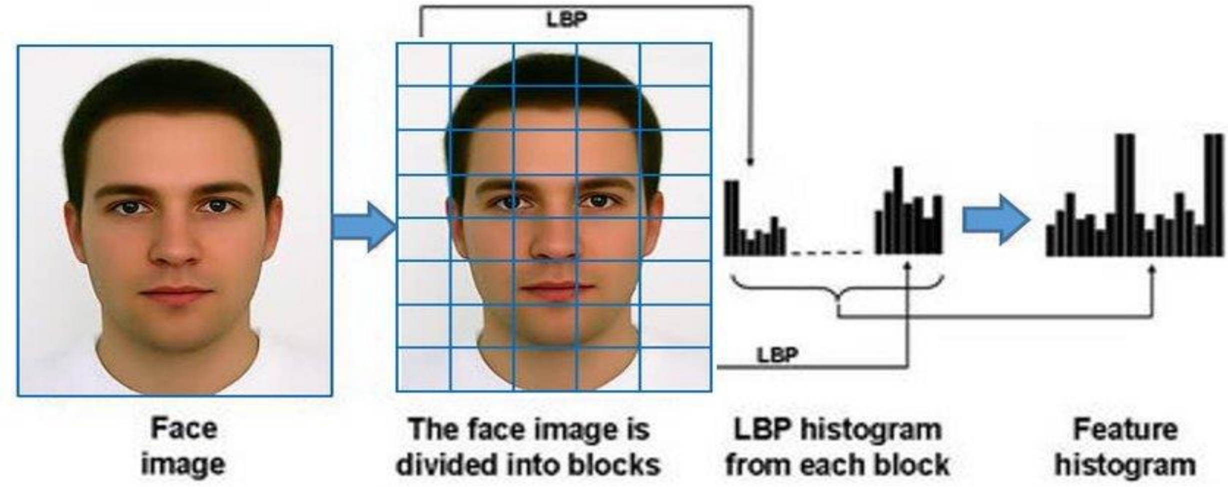

LBP
LBP
Face image
The face image is divided into blocks
LBP histogram from each block
Feature histogram

Figure 2.5: Facial expression extraction with LBP Histogram [23]

2.3.2 Linear Discriminant Analysis (LDA)

LDA is used in pattern classification [24]. It is used to preserve useful features while converting them from high dimensional space to low dimensions. During the process, instead of neglecting useful information in any dimension, LDA develops a new axis by calculating the mean and variance of each data set to combine information from all the data sets on that axis. It makes use of Bayes Theorem to estimate that new dataset contains information from each class (dimension).

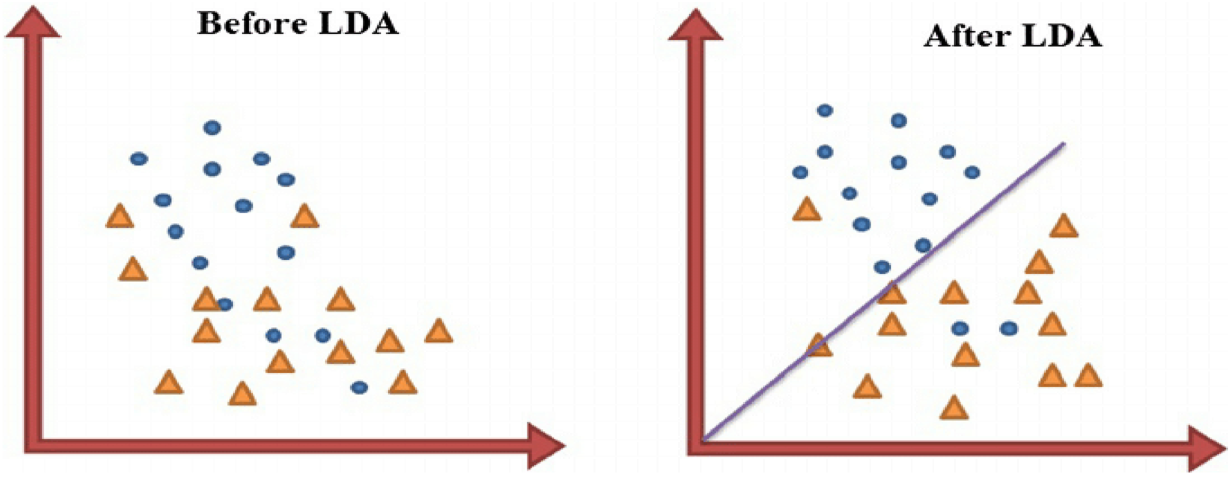


Figure 2.6: Linear Discriminant Analysis (LDA) [24]

2.3.3 Principal Component Analysis (PCA)

PCA unlike LDA is unsupervised algorithm that reduces the dimension by maximizing the correlation between features. Its performance gets affected with illumination, and sometimes causes the algorithm to discard useful information. PCA creates orthogonal axis with the space where there is maximum variance; however, unlike LDA, it doesn't create the separation between features of different classes and often faces issues in discriminating one human face from the other [25].

2.3.4 Fisher Face Method

It is a combination of LDA and PCA. This method works on the principal that instead of looking at all parts of face, the area which shows maximum variation must be selected. It extracts features that separate one individual from the other. It reduces face space dimensions first by using PCA and then provides significant separation of features by applying LDA. By using Fisher Face method, the issues occur in feature extraction due to illumination in PCA are removed [26].

2.3.5 Line Edge Map (LEM)

It is a template-based matching algorithm that creates outlines around the facial features just like pencil made portrait. It just stores the end points of lines thus reducing the memory requirements and calculates a distance to separate each line from the other. After the necessary computation, it saves the template to be used at later stages for comparison [27].

2.3.5 Gabor Wavelet

Gabor Wavelet enhances the face features for example nose, mouth, wrinkles, eyes, scars, etc. for represent face in high dimensional space [28].

2.4 Feature Classification

The final phase of the FER system is the classification that can be realized either by attempting recognition or by interpretation. FER deals with the classification of the face and its features into abstract classes that are entirely based on visual information. Facial expression classification aims to design an appropriate classification mechanism to identify facial expression. Earlier facial expressions were categorized into six basic emotions: Disgust, Anger, Fear, Surprise, Happy and Sad. But after some time, many of the recent research work includes Neutral expression in this list. Hence facial expressions are categorized into seven basic emotions: Disgust, Anger, Fear, Neutral, Surprise, Happy and Sad. [29]. To identify the above listed facial expressions, a process must be able to recognize facial feature movements. According to these different emotions will be classified into seven categories as mentioned in below table 2.1

Table 2.1: Seven Basic Expression Description [29]

Emotion Class	Explanation
Angry	Eyebrows are pulled downward and together. Eyes are wide open, and lips are tightly closed.

Disgust	Eyebrows and eyelids are relaxed. The Upper lip is raised and curled, often asymmetrically
Fear	Eyebrows are raised and pulled together. Eyes are open and tensed.
Happy	Eyebrows are relaxed. The Mouth is open, and Mouth corners are upturned.
Sad	Eyes are slightly closed. Eyebrows are bent upward, and Mouth is relaxed.
Surprise	Eyebrows are raised. Eyes are wide open, and Mouth is open.
Neutral	Eyebrows, Eyes and Mouth, are relaxed.

Support vector Machine (SVM) is one of the most important method for image classification.

2.4.1 Features Classification using Support Vector Machine (SVM)

Support Vector Machine (SVM) is a linear model used to handle classification and regression problems. It works by drawing multiple separation lines between data of two classes and then identifies one line which provides better separation between data among all the lines. This line is termed as generalized line in SVM and the distance between this line and the data set vectors is called hyperplane [30]. This algorithm works on data that is provided as an input and in the result, it provides a generalized line to provide separation between data of different classes.

For example, in the above diagram, yellow line is the output line and considered as generalized because it clearly separates the two data sets, while green line is closer to red data set and if considered as a generalized line will cause features of red data set to dominate over blue data set.

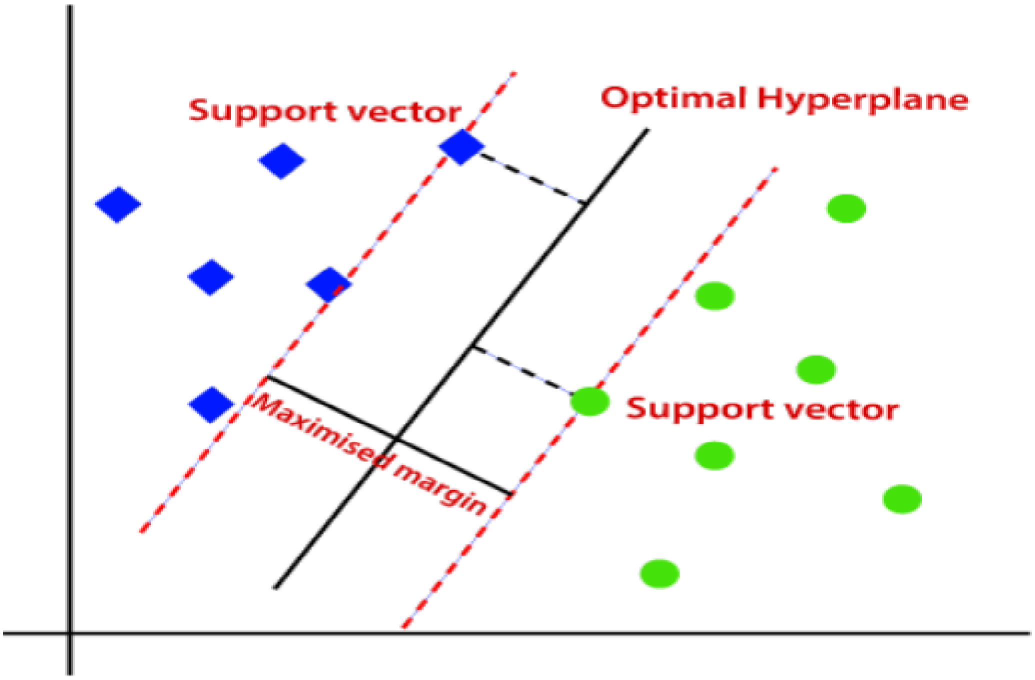


Figure 2.7: Hyperplane and Support vector example in SVM [31]

The next process involves calculating the distance (termed as margin in SVM) between the generalized line and data vectors. In this algorithm the main goal is to select the line which offers maximum distance and separation from data vectors. The maximum margin offers the optimal classification of data as it minimizes the chances of feature overriding. The area of separation for which margin is maximum is called optimal hyperplane. Thus, SVM tries to make a decision boundary in such a way that the separation between the two classes (that street) is as wide as possible.

2.4.2 Features Classification using Artificial Neural Network (ANN*)*

As penetration of digital systems in our daily lives has increased the human-computer interaction, the need for the intelligent systems to properly decode human behavior has led to the development of different intelligent algorithms for the machines to behave in an effective manner in the human-oriented environment. Artificial Neural Network (ANN) systems have been designed to incorporate the decision-making skills in machines similar to that exists in humans. These systems operate using mathematical computations and possess dynamic ability like human brains to make decisions independently. These ANN algorithms consist of several independent decision-making nodes known as neurons. These neurons behave in the same manner as cells operate in the human brain. These algorithms are generally trained at first to decode certain type of information in real-time environment.

One of the robust features of ANN algorithms that make them more suitable for information interpretation (feature extraction) is the assignments of weights. When any neuron learns some attribute during training, it acquires a weight. The trained neuron either passes that weight to neighbors so that they can store the information associated with that weight, thus saving the time for each neuron to learn the same information independently; or the learned neuron holds that weight to itself only, thus allowing other neurons to learn new attributes of the same object independently thus bringing more diversity in the system. The learning process usually accomplished in different layers known as hidden layers between input and output layers. Each hidden layer consists of a set of neurons and after training at the layer; the associated neurons either pass or withhold the learned information for their fellow neurons at the next layer. The process continues till all the neurons in the system learn some attributes. The number of hidden layers can be increased or decreased depending upon the nature of application. More layers allow more comprehensive training and results in more critically designed system [30], [31].

With the development of large databases, high production demands in industries, online gaming, subtle medical treatments etc., the human ability to handle the diverse nature of information recursively offers more challenges in terms of accuracy and quality maintenance. Keeping in view certain factors like human resource constraints for handling large information, human causing errors, and human tiredness have led to the development of the idea of Artificial Intelligence (AI). The reason behind the idea of AI is that the machines should be programmed in a way that can perform large recursive operations intelligently, thus minimizing the chances of errors and eliminating the factors of tiredness and quality deterioration that is associated with natural human behavior. For this purpose, researchers have developed many parameters to map human behavior into digital manipulation units for the machines so that they can not only be programmed at the initial stage but would also be able to behave dynamically in real-time changing scenarios.

2.5 Deep Learning for FER

The concept of AI has evolved in the form of intelligent machines over the years acknowledged as thinking machines in today's era where many sensing technologies and robotics are observed to be working efficiently in human-oriented environment [33]. The incorporation of AI into machines has been challenging for the researchers from the factor of training systems to achieve human-matching intelligence. Many algorithms and datasets have been proposed to improvise these training mechanisms. After several experiments and development of different techniques and algorithms, deep learning has gained success in Machine AI due to its human-brain like working mechanism.

Figure 2.8 explain the artificial intelligence development. The smart intelligence-based systems, that can learn on the dataset, which is provided to it on its own, could be built using machine learning, a subcategory of Artificial Intelligence. Moreover, the subset of machine learning is deep learning, in which deep neural networks could be trained using the machine learning algorithms. New deep neural network is being trained to achieve the better results, if the former network is not giving promising results.

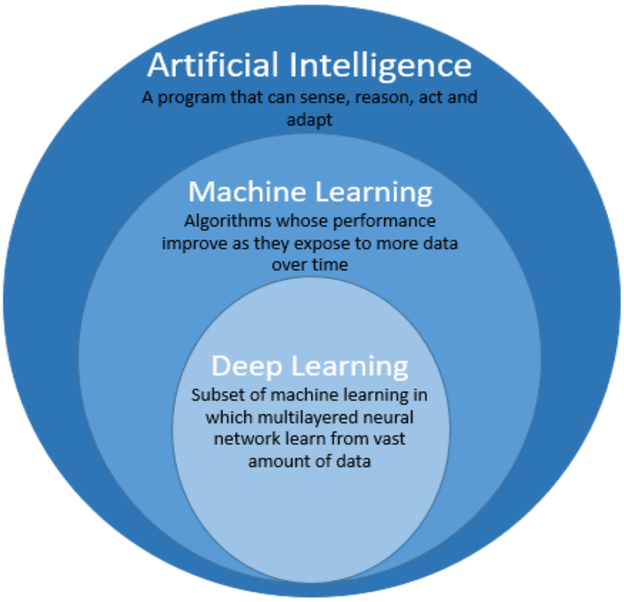

Artificial Intelligence
A program that can sense, reason, act and adapt
Machine Learning
Algorithms whose performance improve as they expose to more data over time
Deep Learning
Subset of machine learning in which multilayered neural network learn from vast amount of data

Figure 2.8: Artificial Intelligence evolution

2.5.1 Machine Learning

Machine learning and artificial intelligence are interlinked and are highly correlated. To create an intelligent system these both technologies are used in connection with each other. Artificial intelligence is study and development of such a machine that is capable of thinking and behaving like human whereas, machine learning is considered to be a subset of artificial intelligence in which system does not need to be programmed explicitly but it learns from input data. Arther Samuel was the first person to describe the term machine learning in 1959. In a traditional software development computer is destined to execute some set of instructions given by the developer of the program whereas, machine learning differs in this context because it focuses on discovering an algorithm that can develop or improve itself on its own instead of being explicitly programmed. Machine learning mainly emphases on decision making or predictions from the given data that is why is also termed as predictive analytics [17]. There are basically three types of machine learning algorithms named as: unsupervised learning, supervised learning and reinforcement learning. Supervised learning makes a model which contains inputs and desired outputs. Inputs given to the system are labeled in this case belonging to particular categories. Regression and classification mainly lie under supervised learning. In contrary unsupervised learning contains only input data without having categorization labels therefore, in this case algorithms learns from unlabeled data and find sequence and structure of data to take decisions. Cluster analysis is an example of unsupervised learning. Third category reinforcement learning focuses on the actions and behavior of an agent in different environments to achieve a goal, based upon which the agent gets rewarded or punished. Chess game is an example of reinforcement learning. This research falls under supervised learning because the data being used for experiments is labeled. Label here corresponds to a unique emotion.

2.5.2 Deep Learning Based Artificial Intelligence

The classical FACS system introduced the rules to classify basic emotions into AUs representing the detection of emotions at conscious level. However, the understanding of deep subtle behavior of humans at unconscious level the knowledge of which is particularly required in psychological treatments, classical FACS have proved less efficient to serve the purpose. With the rising trend of research in the field of computer vision, artificial intelligence (AI) have been introduced to assist in the detection of expressions in unconscious problem solving processes, and consequently have led to the development of deep learning. The schematic of deep neural network is shown in Figure 2.7:

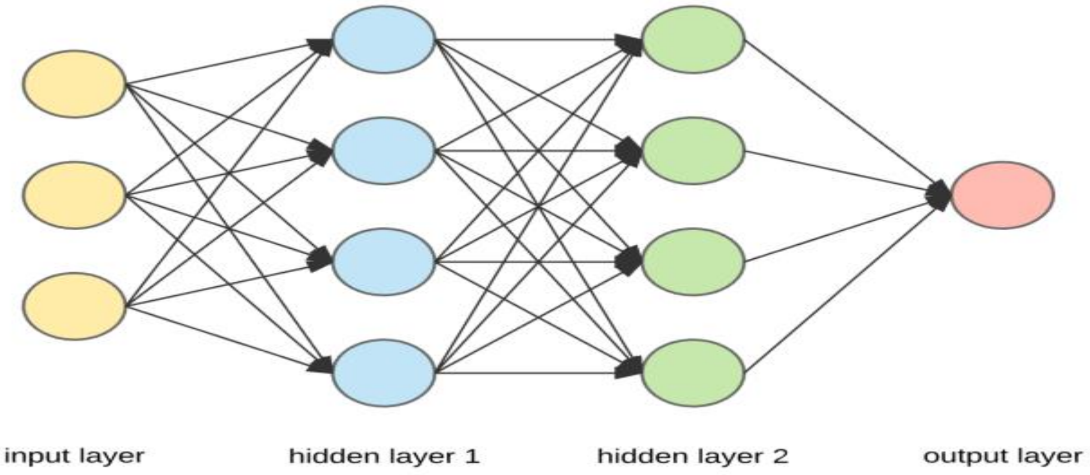


Figure 2.9: Schematic of Deep Neural Network

The Neural network consists of individual units called neurons. Neurons in each layer are connected to neurons of the next layer. When input image is fed to the system, each individual node performs a simple mathematical calculation and transmits its data to all the nodes of next layer with which it is connected.

2.6 Architecture of Convolutional Neural Networks

CNN algorithm operates on different layers following the Neural Network Principle [34]. The different layers of CNN are explained in detail in subsequent sections:

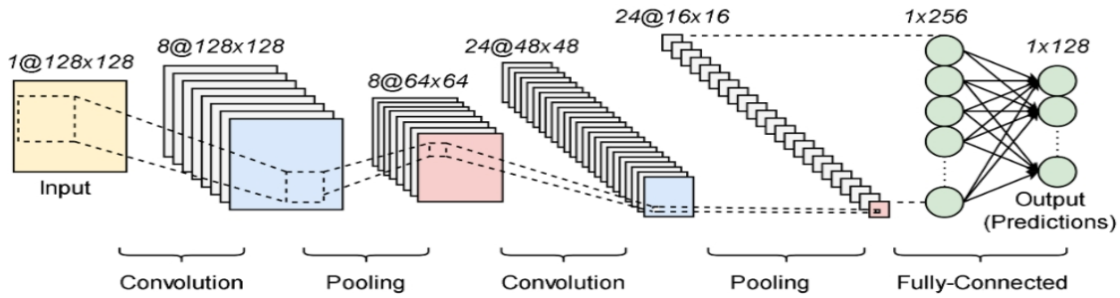


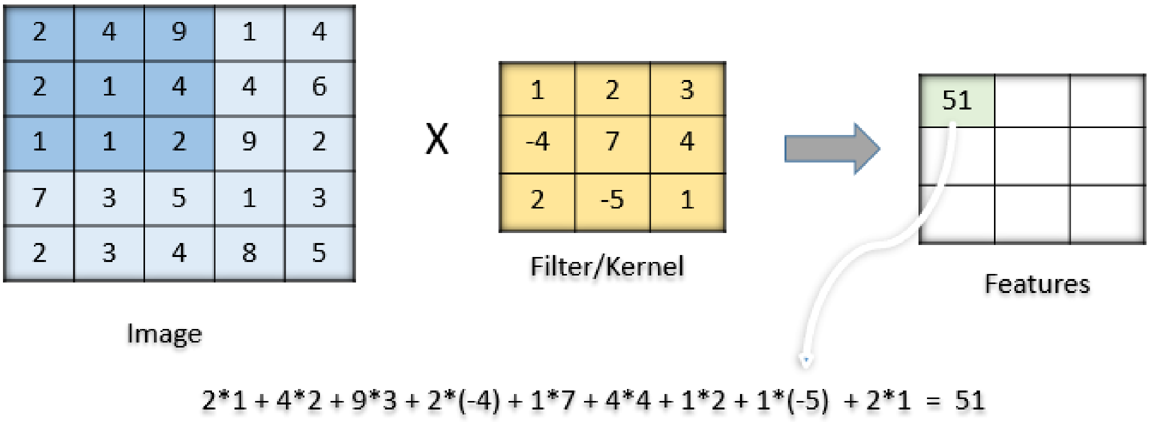
Figure 2.10: Convolution Neural Network Architecture [34]

2.6.1 Convolutional layer

The first layer, known as Convolutional layer, extracts feature from the input image which is usually in the form of a matrix containing pixel values. This layer then extracts most dominant feature known as core or neuron in the form of small matrix. This extracted matrix is also known as a filter. This filter moves along the original input matrix and performs convolutions for each set of pixels, thus resulting in output for each convolution operation [34]. Convolution of input image with filter in term of mathematic is define in equation 2.2:

$$(f * g)(i) = \sum_{j=1}^{m} g(j) \ . f\left(i - j + \frac{m}{2}\right) \qquad \ldots \ldots \ldots \ldots \ldots (2.2)$$

This is the dot product of the input image and the filter. Figure 2.10, 2.11 and 2.12 shown the convolution operation step by step.



2	4	9	1	4
2	1	4	4	6
1	1	2	9	2
7	3	5	1	3
2	3	4	8	5

Image

X

1	2	3
-4	7	4
2	-5	1

Filter/Kernel

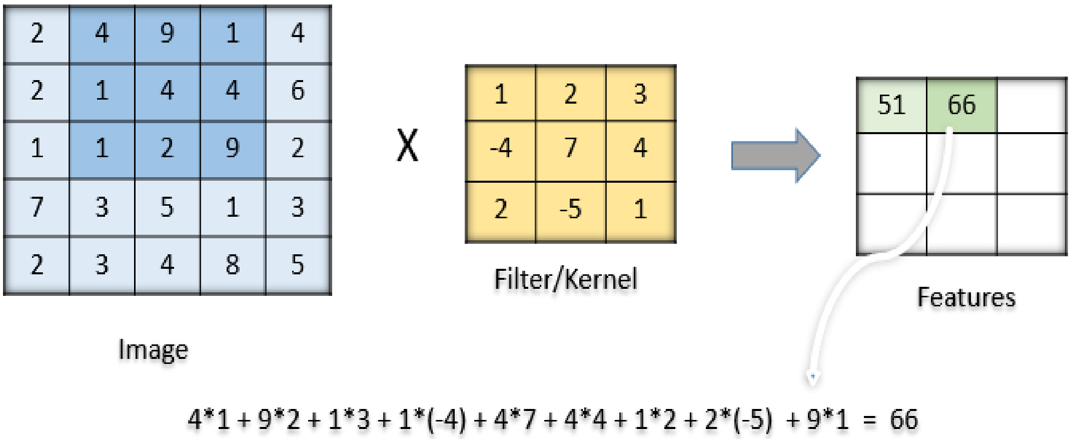
51		

Features

2*1 + 4*2 + 9*3 + 2*(-4) + 1*7 + 4*4 + 1*2 + 1*(-5) + 2*1 = 51

Figure 2.11: Convolution Operation Step – 1

When the image passes through one convolution layer, the result of the first layer becomes the input for the second layer. This filtering operation, in the end, produces a small matrix containing the outcomes of all the convolution operations.



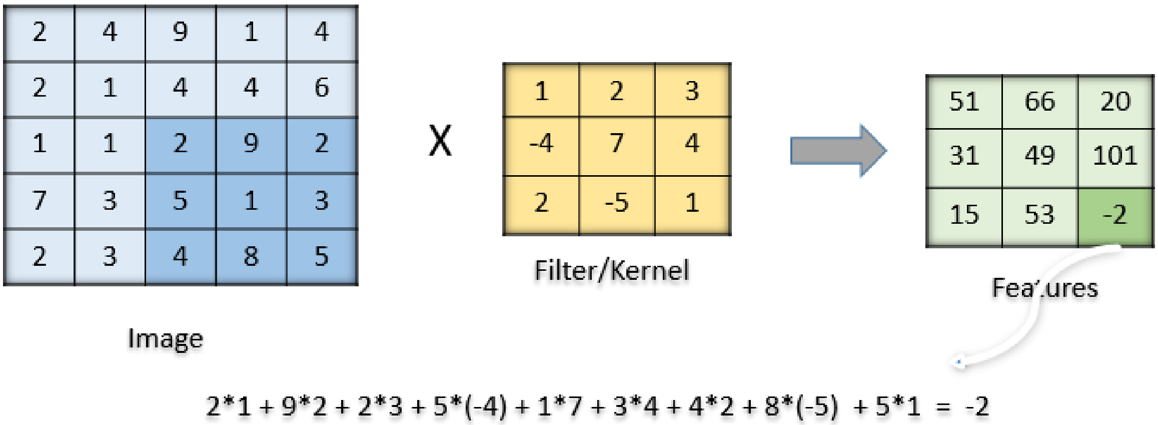
2	4	9	1	4
2	1	4	4	6
1	1	2	9	2
7	3	5	1	3
2	3	4	8	5

1	2	3
-4	7	4
2	-5	1

51	66	

4*1 + 9*2 + 1*3 + 1*(-4) + 4*7 + 4*4 + 1*2 + 2*(-5) + 9*1 = 66

Figure 2.12: Convolution Operation Step – 2



2	4	9	1	4
2	1	4	4	6
1	1	2	9	2
7	3	5	1	3
2	3	4	8	5

1	2	3
-4	7	4
2	-5	1

51	66	20
31	49	101
15	53	-2

2*1 + 9*2 + 2*3 + 5*(-4) + 1*7 + 3*4 + 4*2 + 8*(-5) + 5*1 = -2

Figure 2.13: Final step for feature extraction

Convolution of facial expressions image with multiple filters can achieve operations such as blur, edge detection and filter applying for sharpen the image.

2.6.2 Receptive Field

The size of the region in the input that produces the feature is called receptive field in CNN. Basically, it is a measure of association of an output feature (of any layer) to the input region.

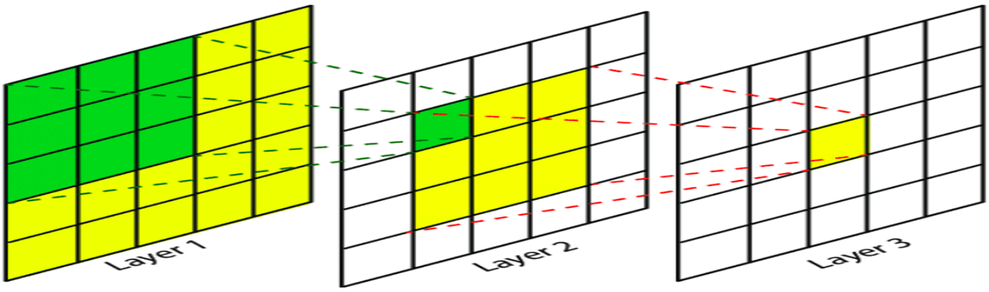


Figure 2.14: Receptive Field operation [35]

Large objects could be unable to recognize by the small receptive field in object detection. In order to overcome these problems some multi scales methods are used. Also, for capturing the large motions in motion-based tasks such as optical flow estimation and video prediction, an adequate receptive field is required.

For instance, in the figure 2.15 below a car is shown. In the image, two receptive fields: the green and the yellow one is seen to be serving the purpose.

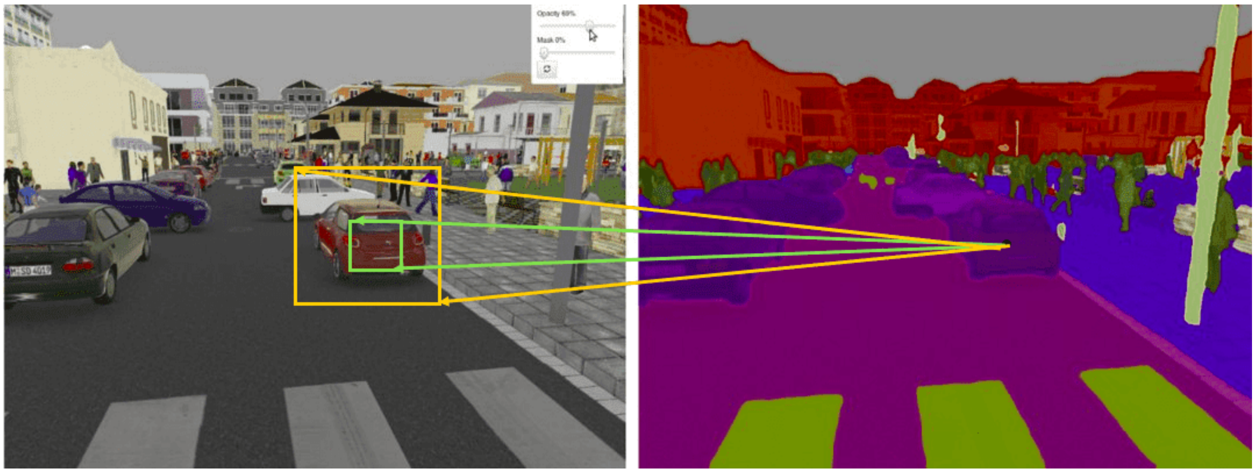


Figure 2.15: Receptive Field for feature extraction [35]

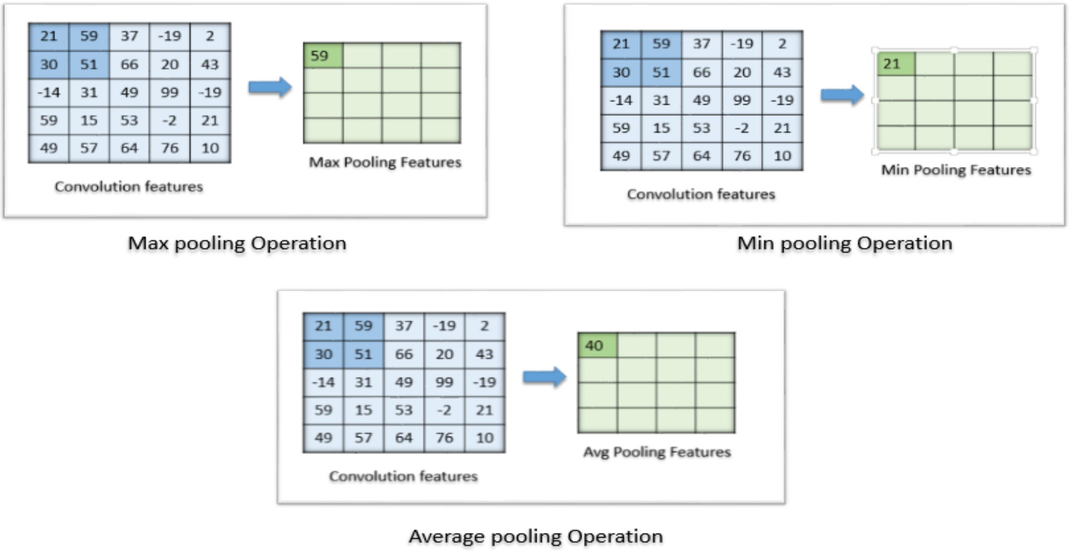
2.6.3 Weight Sharing

Weight sharing emphasis the fact that two or more layers in the network will use the same weights i.e., the same parameters will be used to represent different transformations in the system. When one network trains, it will update those weights. The second network will then use those same weights and so on.

For instance, in the image containing multiple cars, if the network doesn't use weight sharing, the network will have to independently learn the appearance of cars at every distinct image location. This is because a filter that is allocated for the bottom left part of the image will only see bottom-left image patches. Thus, to learn the function well, the training data will also have to contain images with cars (in various poses/models) at every image location, which in turn requires larger training set. On the other hand, by sharing weights, the network will be able to learn a single filter for cars no matter where the car appears in the image.

2.6.4 Pooling Layer

The pooling layer then works on the width and height of the image. This layer down samples the output matrix or merges all the convolutional matrices by removing all the over-riding information, thus resulting in a compressed image [36]. It is similar to the convolution layer but it takes the max of the region from the input overlapped by the filter.



21	59	37	-19	2
30	51	66	20	43
-14	31	49	99	-19
59	15	53	-2	21
49	57	64	76	10

Convolution features

59			

Max Pooling Features

Max pooling Operation

21	59	37	-19	2
30	51	66	20	43
-14	31	49	99	-19
59	15	53	-2	21
49	57	64	76	10

Convolution features

21			

Min Pooling Features

Min pooling Operation

21	59	37	-19	2
30	51	66	20	43
-14	31	49	99	-19
59	15	53	-2	21
49	57	64	76	10

Convolution features

40			

Avg Pooling Features

Average pooling Operation

Figure 2.16: Pooling Layer Convolution Operation

The convolution and pooling operations occur at each hidden layer between input and output layer to allow neural network to break down complete function into specific transformations of the data. Each hidden layer function produces a defined output and helps to learn different aspects about the data by minimizing an error.

2.6.5 Drop Out

The process of setting values of some input neurons to 0 is called “drop out” in CNN. Weights in CNN are used to represent some specific features acquired by neurons in the training process. . If we make all the neighbors (neurons) to rely on those specific weights only, the model becomes too fragile to detect multiple variations and leads to complex coadoptions.

However, if we allow the chance to all the neighbors (neurons) in the network to participate and make predictions by asking some neurons (weighted/ trained) to stay quiet for some time, new independent internal representations will be learned by the network. As a consequence, the network will become less sensitive to the specific weights of neurons. This in turn results in a network that is capable of better generalization and is less likely to overfit the training data.

2.6.6 Batch Normalization

Batch Normalization works with Drop Out layer and allows every layer in the network to learn independently. It regularized the data learning in between the layers [37].

2.6.7 ReLU Layer

Using of non-linearity is crucial in image processing as linear functions only allows for simple multiplication and addition due to which nothing interesting information can be extracted from them. ReLU is a non-linear function like sigmoid and tanh. However, it provides benefits of low computational complexity, no saturation at higher values and faster convergence. It outputs 0 for negative values and outputs the same value as given in input (if positive).

2.6.8 Flattening Layer

All the convolution and pooling operations result in rectangular matrices. But as the rectangular matrices cannot directly input information for classification, the Flattening layer plays its role in converting the N dimensional matrix of data into a 1-dimensional array for inputting it to the next layer, thus resulting in a fully connected dense layer [37].

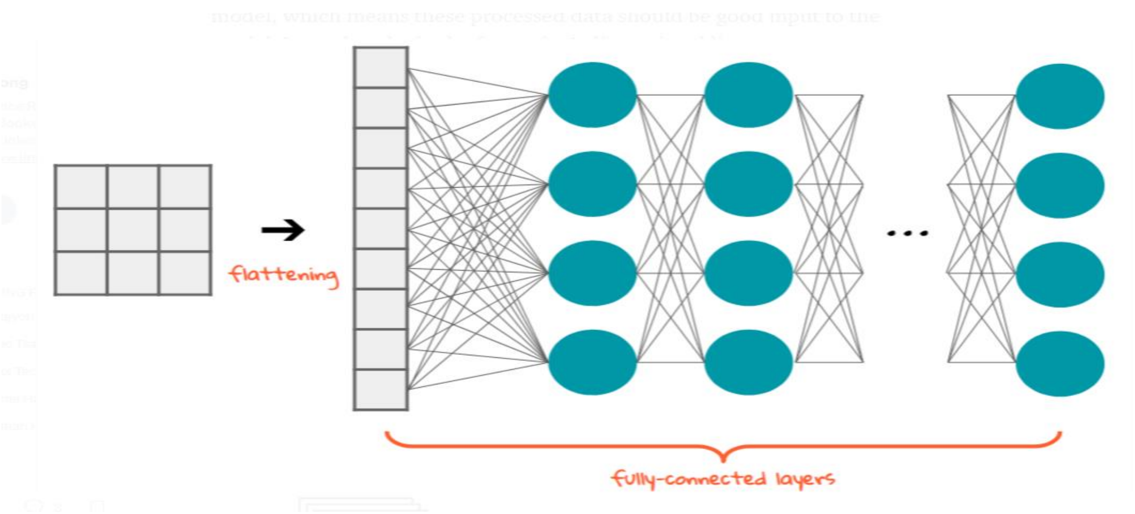


Figure 2.17: Flattening Layer Operation [37]

Chapter#3

Literature Review

CHAPTER 3: LITERATURE REVIEW

The research in this thesis involves those systems that operate by evaluating the human expressions from their face using deep learning algorithms. This chapter highlights all the conventional and latest techniques that have been used for facial expression recognition from images. The foundation of AI system for operation is the accurate interpretation of human environment. The incorporation of AI and its usage by the machine depends upon the application. The conventional approaches for FER detection from images involved manual marking of each face area where the minor misalignment due to motion, and changes in lightening, etc., could produce the erroneous results. The experts had to carefully pre-process the images before feature extraction resulting in much consumption of time with little probability of accuracy. On the other hand, deep learning techniques, due to their dynamic ability of self-learning, have overcome such challenges. Besides reducing head of pre-processing of images, deep learning algorithms offer advantages of auto-adjustment to changing factors like illumination and occlusion. Moreover, they are also able to handle large amount of data. Despite the numerous advantages of deep learning algorithms, some drawbacks are also associated with them as they require large amount of capacity for processing at different layers. In addition, the layer-to-layer propagation of data sometimes results in loss of important information at intermediate layers which require appropriate selection of pooling and debuffing. Researchers are still working on developing techniques to improvise the processing capability and accuracy of existing deep learning algorithms.

3.1 Conventional FER Approaches

Conventional FER techniques mainly involve the large overhead of image pre-processing due to their intrinsic property of detecting manually marked areas in images which requires a meticulously marking of the geometric features, or appearance features, or a hybrid feature on the target face and additional procedures of feature extraction and feature classification [38].

3.1.1: Image Pre-processing

Irrelevant information in the images at this stage is filtered out so that the detection of the desired object or expression can be achieved with better accuracy. This step directly impacts the feature extraction capability of the system and affects the classification process. The images having complex backgrounds pose more challenges in terms of feature extraction and the critically discriminative marking of different objects is the essential factor for the good performance of the detecting system [39]. Different techniques and methods are used for image pre-processing in FER algorithms. The first step in the image pre-processing is the removal of noise by applying different filters that include but are not limited to Average Filter (AF), Gaussian Filter (GF), Median Filter (MF), Adaptive Median Filter (AMF) and Bilateral Filter (BF). The next important step is the face detection as its detection is affected by certain factors like presence of shading objects, changes in color of skin, rapid movement in face area, etc. which requires normalization of the color scales and the proper orientation of detection plane to reduce the complexity of expression detection while ensuring the accurate recognition of features [39].

According to [40], Papageorgiou et al. in 1998 introduced a framework that makes use of frequency based haar wavelet method for image detection. Later in 2001, Viola and Jones made use of frequency spectrum to define changes in skin color and edges of facial features following the principle of Haar-Wavelet technique. In [41], the researchers used this technique to improvise the computation process by cascading different Haar classifiers together. In [42], Weilong Chen et al. worked on the factor of changing light conditions and introduced the normalization technique to reduce the impact of varying illumination on the expression detection. They considered using Discrete Cosine Transform (DCT) to detect faces under varying illumination conditions. Their technique worked well for illumination changes but failed to produce desired results under shadowing conditions. In [43], Owusu et al. introduced the technique of limiting the size of image while preserving the useful information using the Bessel down-sampling approach. Biswas et al. in [44] enhances the smoothness of input images by resizing them using the Gaussian filter.

In [45], Idrissi et al. used the median filter during normalization to enhance the quality of input image. Zhang et al. in [46] and Happy et al. in [47] introduced a localized method for image pre-processing following the principle of Viola-Jones to help in the detection of facial images from the input image. Adaboost and Haar algorithms mainly help in the detection of size and area of face in the images while face alignment is usually accomplished using the SIFT (Scale Invariant Feature Transform) algorithm. ROI (Region of Interest), on the other hand, regulates the face dimensions by dividing the color components and facial features.

3.1.2: Feature Extraction

The next important step is the extraction of facial features so that those extracted features can be processed to produce desired results. This stage is considered crucial in facial expression recognition as it leads to classification of data on the basis of which machines make their decisions regarding a particular action. Most commonly geometric features extraction, appearance-based features extraction or hybrid features extraction techniques are used to save the desired information.

In [48], Happy et al created histograms to represent different face features using a technique known as Local Binary Pattern (LBP) and then make their classifications using Principal Component Analysis (PCA). Ghimire et al. in [49] retrieved appearance-based features by marking the face features into different regions known as domain-specific local regions. The important regions are then extracted using an incremental search approach which focuses on the specific region and helps to improve the recognition accuracy of FER algorithm. Ghimire and Lee in [50] made use of two different geometric features relying on the position and angle of 52 facial points. In the first stage, the distance and angle between different facial points is calculated and in the next stage, these calculated parameters are subtracted from the corresponding distance and angles in the frame of the video sequence. This technique made use of two classifiers i.e., AdaBoost and SVM for classification process. Aruna Bhadu et al. in [51] used a hybrid feature extraction approach by combining Discrete Cosine Transform and Wavelet Transform, and then applied Adaboost as a classifier.

Some researchers are working to design systems that will operate on Infrared images because images in visible light shows more variation due to changes in illumination and makes the detection of the features difficult for the machine. Following this concept, Zhao et al. in [52] worked on Near-Infrared (NIR) video images. This study uses component-based facial features to combine geometric and appearance information of the face. For FER, an SVM and sparse representation classifiers are used.

In [53], Shen et al. used a technique of calculating temperature difference between different sub-regions of face to identify features, and then the researcher applied Adaboost for classification of features. Szwoch and Pieniazek [54] used Microsoft Kinect sensor's depth channel to recognize facial expressions and emotions without using camera. They analyzed the local movements of facial area to recognize facial expressions and map relations of those expressions with particular emotions. Wei et al. [55] used Kinect sensor to identify color and depth information of the features. The researchers made use of face tracking algorithm to extract features, and then used random forest algorithm for their classification.

3.1.3: Facial Expression Classification

The last stage of Facial Expression Recognition is the classification of extracted features where different classifiers are used to create relations between extracted features and emotions. Formerly, the human expressions were classified into six basic emotions that included happy, sad, fear, anger, disgust and surprise, and for many years researchers worked on these six expressions to accurately identify them and deduce other complex emotions from the manipulation of these expressions. However, with the passage of time researchers felt the need to introduce another expression i.e., Neutral to complete the expression set because most of the time the movement of brows and eyes doesn't reflect any emotion other than Neutral. Many classifiers have been developed so far to accomplish the task of classification that include but not limited to KNN (K-Nearest Neighbor), SVM (Support Vector Machine), Naïve Bayes Classifier, AdaBoost Classifier, HMM (Hidden Markov Model), Decision Tree and NN (Neural Network).

According to research work in [56-58], KNN classifier has gained popularity due to its simplicity and minimum computational requirements. Besides the advantages of this classifier, the major drawback that is associated with this classifier is that it is sensitive to type of input data. Its results vary by varying size of neighbourhood i.e., value of k, and no technique has been developed so far to decide the optimal size of neighbourhood (value k). Most of the time the value of k is dependent only upon the nature of application, and therefore the precise optimization of the output results is not possible.

.

SVM Classifiers are well known classifiers to obtain generalization for complex data. They also make use of kernel functions to convert indivisible data into linear separable samples thus enabling the classifier to process high dimensional data. Liyuan Chen et al. in [59] proposed a research work to recognize relevant expressions from irrelevant expressions using SVM. The researcher used both linear function and Radial Basis function (RBF) for classification purpose, and as a result 80% accuracy was achieved for the cases used for the study under research work. Many researchers [60-62] have used SVM classifier in their proposed approach for classification purpose.

With the incorporation of Neural Networks in FER recognition, many researchers are now using these classifiers in their studies. The variants of neural networks have been developed over the years with different improvements and variations in classification patterns. The reason behind the development of different variants is the number of hidden layers and learning mechanism. Each application requires a certain number of hidden layers for producing the desired results for instance, a single hidden layer ANN cannot perform XOR problem that requires an additional layer to be incorporated in the system. Multilayer Feed Forward Neural Network (MFFNN), a variant of ANN, using back propagation is the algorithm in which the back propagation technique helps in learning of the neurons on different layers where weights are iteratively updated after the learning of each tuple. It should be noted that in feed forward system the information learned at any stage cannot be shared and forwarded to the previous hidden layers or input layer. The data moves only in one direction i.e., from input to output. The back propagation algorithm only helps in replacing the weights of neurons as the new information is learned by the other neurons in the network. The Bayesian neural network classifier acknowledged as stochastic model is also the type of neural networks. This classifier makes use of probabilities to produce desired results. In [63], the researcher used Bayesian classifier with back propagation algorithm to test the accuracy of different case studies. Some researchers [64-65] have used a probabilistic neural network as a classifier in facial expression recognition.

The conventional techniques of feature extraction and classification are considered useful due to their low computational complexity and less hardware dependency, and they are still considered useful in designing of real-time embedded systems. However, due to large complexity of marking recognition area manually in conventional techniques, deep learning method is considered better for complex applications where the classifier can automatically adjust to the slight variations. As no model has been developed so far which can address all the issues of conventional and deep learning methods; therefore, all the existing models are in use, and researchers usually modify the existing models according to their needs depending upon the nature of application.

3.2 Deep Learning-based FER Approaches

Deep-learning has gained popularity in recent years due to its capability to auto-learn the new features based on the stored information during the training process, thus minimizing the need to train the system over and over again for new changes. The intrinsic intelligence of the system makes it efficient to perform well in real-time environments where the rapid movements, change of illuminations, shading, and non-discriminative object identification are the major problems. Moreover, deep-learning algorithms also possess the capability to handle large amount of data as manual pre-processing is not involved in these algorithms. Deep-learning algorithm consists of two variants known as convolutional neural network (CNN) and recurrent neural network (RNN). The major drawback associated with these techniques is the involvement of complex hardware as different layers are used to process the data; nevertheless, in CNN, researchers have tried to address this issue by enabling "end-to-end" learning directly from input images, thus reducing the needs of pre-processing.

In [66], Lopes et al. investigated the impact of data-preprocessing before training the network. The researcher made some changes like rotation correction, cropping, down-sampling with 32x32 pixels and intensity normalization before feeding the data to two-layered CNN. The databases used for the research work are CK+, JAFFE, BU-3DFE. According to the test results, pre-processing the images before feeding them separately into the neural network helps in achieving better classification as they directly impact the learning process of the model.

Agrawal et al. in [67] studied the impact of the variation of the CNN parameters on the recognition rate. The database used during the research is FER2013. The researcher used varying size and number of filters and optimizers that include Adam, SGD, and AdaDelta on a two layered CNN. In this work, the researchers produced two models of CNN having accuracy of 65.23% and 65.77% respectively. Deepak jain et al. in [68] improved the cropping and normalization intensity of images by using a deep CNN model having two residual blocks, each with four convolutional layers. The databases used during the research are JAFFE and CK+.

Liu et al. in [69] introduced a Boosted Deep Belief Network (BDBN) comprising three stages. In the proposed model, the first stage selects the first frame with neutral expression and then last three frames are selected from each image to obtain samples for classification. The researchers used CK+ database for the case study. According to the test results, the proposed framework achieved dramatic improvements over current state-of-the-art algorithms.

Burkert et al. in [70] introduced a new CNN architecture having four components. The images are firstly pre-processed using a convolutional layer, and then downsampled in the second layer. The researchers introduced a new block known as FeatEx, inspired by the GoogleNet. This block is considered as the fundamental structure in this architecture. After two concatenated FeatEx blocks, the extracted features are fed into a fully connected layer to perform the classification. The experiment on the CK+ dataset achieved a recognition rate of 99.6%.

3.2.1 Deep Learning Approaches Using FER2013 dataset

Khaireddin et al. [71], researchers have emphasized the fact that optimizing the learning behaviour of the ER model is crucial in the development of reliable system with maximum accuracy. In the case study the researchers first did tuning of the system using SGD where optimal batch size and the best drop-out rate was determined by performing grid search to find the hyper parameters to control the learning process. In the study, the researchers performed the experiment with SGD, SGD with Nesterov Momentum, Average SGD, Adam, Adam with AMSGrad, Adadelta, and Adagrad on dataset FER2013 using a VGGNet to find the best optimizer in training the ER Model. The experiment was conducted in two phases. In the first phase of experiment, fixed learning rate of 0.001 was used to test the model. However, in the second phase, a learning rate scheduler was used with a starting learning rate of 0.01 and a reducing factor of 0.75 for 5 epochs as a result of which the SGD with Nesterov momentum was found to provide accuracy of 73.2 % and 73.5 %respectively. In the third phase of experiment, the researchers determined the optimal learning rate scheduler by performing experiment with 5 different schedulers: Reduce Learning Rate on Plateau (RLRP), Cosine Annealing (Cosine), Cosine Annealing with Warm Restarts (CosineWR), One Cycle Learning Rate (OneCycleLR), and Step Learning Rate (StepLR) as a result of which Reducing Learning Rate on Plateau (RLRP) was observed to achieve a validation accuracy of 73.59 % and a testing accuracy of 73.06 %.

Qihua Xu et al. [72] escalated the idea that effective FER model design can be made by combining different CNN networks or merge multiple CNN with attention mechanism, GCN, GAN, and other network types. In this study, the researchers investigated attention mechanism, multi-network transfer learning, and decision-level feature fusion technology using a large-scale face dataset C-MS-Celeb and two static facial expression image datasets RAFDB and FER2013+ as experimental data and five FER Models i.e., resnet18, vgg16, squeezenet v1.1, densenet121, and resenet50 to analyses the accuracy of FER. The results of experiments showed that the models using transfer learning offers more recognition accuracy. The vgg16 and resnet18 with transfer learning appeared to achieve the better performance with accuracy of 86.8% and 87.82%. Moreover, it was found that by adding the attention module, the recognition accuracy further increases by 2.27%, thus indicating that using low recognition weights suppresses the contribution of low-quality images in the training phase. In the multi-feature fusion module, tuning the proportion parameter to 0.7 on the RAF-DB and FER2013 + datasets, the fusion result exhibited the best performance in hard-recognition features, and the final recognition accuracy on the two datasets was found to be 88.23% and 89.51%, respectively.

Jun Liu et al. [73] proposed an end-to-end deep model has been presented by researchers with data enhancement method including a new hybrid feature representation, and an effective classification network to improve the face recognition rate. During the study, the researchers have used three benchmark datasets including the AR face dataset, the FER2013 database, and the CK+ dataset to compare the performance of the proposed model that is found to be providing 98.6%, 94.5%, and 97.2% accuracy respectively over the existing state-of-art work.

Shiqian Li et al. [74] highlighted the fact that general image classification networks (e.g., VGG, GoogLeNet) leads to inadaptability while applying to some Facial Expression Recognition (FER) tasks and also require large parameter size. To overcome such issues the researchers introduced the use of light-weight Facial Expression Recognition Network Auto-FERNet along with effective relabeling method, Facial Expression Similarity (FES), to remove the uncertainty problem caused by environmental factors and the subjectivity of annotators. The research results showed an accuracy of 73.78% on FER2013 without ensemble or extra training data.

A deep-learning-based scheme [75] has been introduced having two-branch deep convolution neural network (DCNN) model to improvise the extraction of detailed edge information of image over the existing single-branch VGG-16 and VGG-19 models. Instead of just exploring geometric features, such as edges, curves, and lines by a first branch, the holistic features can also be extracted by using the second branch.

Table 3.1: Literature Review of FER2013 and CK+ datasets

Year	Author	Technique	Database	Accuracy (ACC) (%)
2013	Yichuan Tang (FER2013 Winner) [77]	DLSVM (Linear Support Vector Machines)	FER2013	71.20%
2021	Khaireddin et al [71]	VGGNet with different optimizer	FER2013	73.06 %.

2021	Qihua Xu et al [72]	Merge multiple CNN with attention mechanism (Result shown with vgg16 & resnet18)	FER2013 RAF-DB	89.51%, 88.23%
2021	Jun Liu et al [73]	Combine model (VGG & Resnet)	FER2013 AR Face CK+	94.5% 98.6% 97.2%
2021	Shiqian Li et al [74]	Light-weight FER Network (Auto-FERNet along with effective relabeling method)	FER2013	73.78%
2021	Karnati Mohan et al [75]	Two-branch deep convolution neural network (DCNN) VGG-16 and VGG-19	FER2013 JAFFE CK+	77.10% 95.63% 97.99%
2021	Luan Pham et al [76]	ResMaskingNet	FER2013 VEMO	74.14% 65.94%
2020	Huanhuan Ran, Shiping Wen et al [78]	DCNN with different activation function (ReLU, Leaky ReLU, Tanh)	FER2013	89.92%, 90.30% 90.43%
2020	Behzad Hasani et al [79]	BReG-NeXt-50 (With only 3.1M training parameters and 15 MFLOPs,)	Fer2013 AffectNet	71.53% 68.50%
2020	Prateek Chhikara et al [80]	CNN-SVM (CNN with SVM as an additional classifier)	FER2013	71.64%
2019	Wentao Huaat et al [81]	Three CNNs-based models (VGG16, Resnet50 & Resnet101) with different structures as the sub-networks	FER2013 JAFFE AffectNet	71.9% 96.44% 62.11%
2019	Trinh Thi Doan Pham et al [82]	VGG-Face with MPL model	FER2013	69.18%
2019	Trinh Thi Doan Pham	DenseNet with MPL model		71.02%

	et al [82]			
2019	Nguyen et al [83]	Ensemble of Multi-level Convolutional Neural Networks	FER2013 AFEW 7.0	74.09% 49.3%
2018	Ming Li et al [84]	Deep-learned Tandem Facial Expression	FER2013	ResNet18 : 83.1% TFE-joint learning 84.3%
2018	Guohang Zeng et al [85]	HoloNet with SoftMax	FER2013 CK+ JAFFE	61.86% 97.35% 83.57%
2018	Soodamani Ramalingam et al [86]	VGG16 & VGG19	FER2013	78%
2017	Jia Xiang et all [87]	MTCNN (Multi-task Cascaded Convolutional Network)	FER2013	60.7%
2017	Sarasi Kankanamge et el [88]	LSTM Networks	FER2013	71.5%
2017	Zhi Li [89]	Discriminative CNN	FER2013 CK+ JAFFE	85.20% 95.21% 97.65%
2017	Philip Lu et al [90]	VGG13 Networkk	CK+	95.70%
2016	Ali Mollahosseini et al [91]	CNN model (2 Convolutional layers with max pooling and then 4 Inception layers.)	FER2013 CK+ MMI	66.4% 93.2% 77.9%
2016	Yuchi Huang et al [92]	D-CNN + Hypergraph	FER2013	77.4% 98.6%
2015	Hong-Wei Ng et al [93]	CNN Network with fine tuning and transfer learning	Fer2013	55.6%

3.3 Limitations and Gaps

It is important to note that many factor can adversely affect the performance of the FER system including occlusion, noise and illumination change and most significantly, facial expression resulting from varying conditions in different region and different emotional states among various individuals. In real world images, extracting features becomes more challenging than in controlled environment.

According to our literature review detail, it can be concluded that researcher is trying to figure out a model that can predict facial expression in real-time from images taken in the real world so that they can improve the performance of existing CNN models. In this study, deep learning techniques are used to improve facial expression recognition accuracy by using images to recognize facial expression. We investigated different method of deep learning to improve the accuracy of facial expression based on less resolution images. To extract essential features from images and improve FER accuracy for real-time expression datasets containing a variety of face expression, we proposed an efficient Convolution neutral network architecture. In addition, this work presents an effective model that achieves outstanding performance over existing state-of-the art technologies.

Chapter#4

Experimental Methodology

CHAPTER 4: EXPERIMENTAL METHODOLOGY

Our technology for recognizing facial expressions is described in this chapter. We discussed techniques used for extracting distinct features from images and methods for classifying based on those features. In this chapter all the details of the proposed methodology is explained to accomplish the task of real-time FER in classifying one of the basic facial expressions. We have discussed FER2013 dataset, CK+ dataset and JAFFE dataset in detail in section 4.1. The details for classification of images are described in Section 4.2. Finally, we present the adopted modify method and the latest CNN-based deep learning method in the upcoming Sections 4.3 and 4.4, respectively.

4.1 Dataset

Emotion Recognition from Facial Images, being the popular subject of research, have attracted ample number of researchers in the last few years due to which many benchmark datasets are publicly available to evaluate the FER approaches. Some datasets are capable for emotion recognition under controlled lab environments and hence poses challenges in the real-world scenarios, while other datasets work well in real time environment offering recognition in images with multifarious variations. An evaluation of the proposed model is performed using real-time controlled FER2013 dataset and uncontrolled CK+ dataset by changing the parameter of the network such as batch size, learning rate and epochs.

4.1.1 FER2013 Dataset

The FER2013, being a huge (greater than 35,000 images) dataset, offers a various range of variations in images like skin color difference, blockage, Person age, face position and quality of the image of face, and its samples have been extracted from real-world Internet data. FER2013 is an approved dataset which has been handled in ICML competitions as well as in many research works. It is one of the most challenging datasets, which can easily accessible via Kaggle website [94] containing the dataset of 35,887 grayscale images standardized to 48x48 pixels. The dataset contains a large collection of images, it exhibits only 7 facial expressions images viz. distributions of (i) Angry with 4,953 images, (ii) Disgust with 547 images, (iii) Fear with 5,121 images, (iv) Happy with 8,989 images, (v) Sad with 6,077 images, (vi) Surprise with 4,002 images and (vi) Neutral with 6,198 images, which make it not a well-adjusted and composed dataset. A few samples taken from the dataset randomly are shown in Figure 4.1

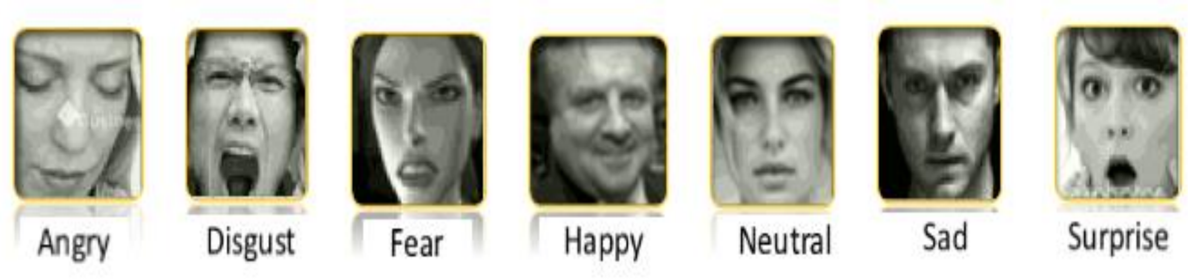
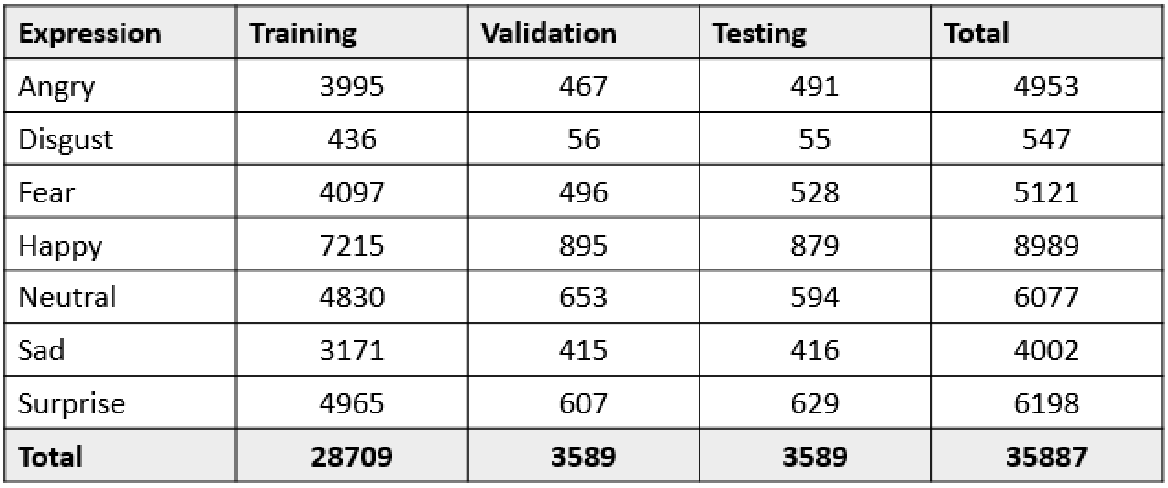


Figure 4.1: FER2013 Dataset Sample

FER2013 dataset mainly comprises of 35887 facial expression images, that includes 28709 training sample images, 3589 images for public verification, and 3589 are for the verification purpose. Each image is scaled to size 48×48 and have a grayscale composition. To create this database and to register the faces automatically, we made use of Google search API. In comparison to other datasets, FER offers more variations in the images, including partial faces, facial occlusion (mostly with a hand), eyeglasses and low-contrast images. However, this dataset also provides an excel file; saves facial expressions, images related data, and purpose data to a csv file and does not support provision of direct image extraction. Table 4.1 explain the complete distribution of the FER2013 facial images.

Table 4.1: Distribution of the FER2013 facial images



Expression	Training	Validation	Testing	Total
Angry	3995	467	491	4953
Disgust	436	56	55	547
Fear	4097	496	528	5121
Happy	7215	895	879	8989
Neutral	4830	653	594	6077
Sad	3171	415	416	4002
Surprise	4965	607	629	6198
Total	**28709**	**3589**	**3589**	**35887**

4.1.2 Extended Cohn-Kanade (CK+) Dataset

Cohn-Kanade Association research on automatic facial image analysis, conducted on coded facial expression database. The CK+ dataset [95] is one of the largest and most comprehensive used laboratory-controlled facial expression classifications exist. CK+ dataset is used to analyze the performance measures of the proposed model in real situations. This dataset consists of image sequences 593 in number from 123 subjects, 327 of which are labeled with one of the seven basic facial expressions. By using two hardware synchronized Panasonic AG-7500 cameras facial behavior of 210 adults was recorded. The age of participants was ranging between 18 to 50 years, which includes 69% female, 13% Afro-American, 81%, Euro-American and 6% other groups. The experimenter instructed the participants to perform a series of 23 facial displays, which included combinational and single action units. Each display began and ended in a neutral face with any exceptions noted. The image consists of 9-60 frames for each facial expression, while the participants were directly facing the camera. Some sample of CK+ dataset is shown in Figure 4.2.

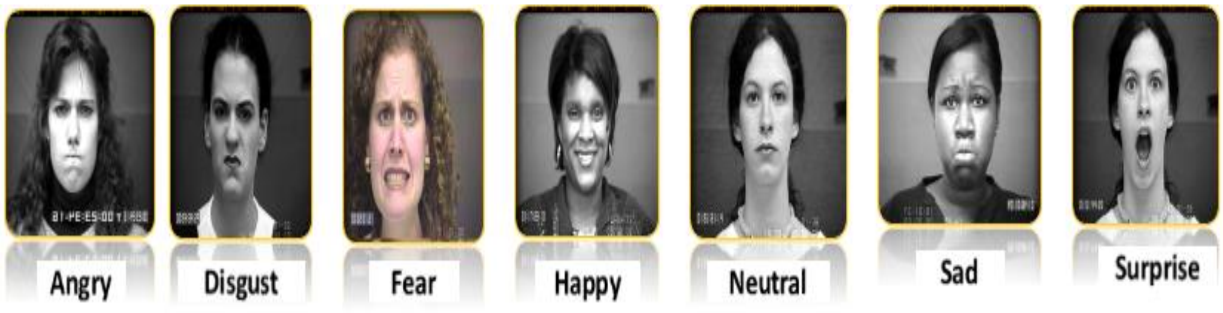


Figure 4.2: CK+ Dataset Sample

The resolution of each image was scaled to 640×490 640×490 pixels or 640×480 640×480 pixels and was in grayscale or in color. In this work, from 327 labeled image sequences we have extracted the last three frames from each, which is made up by a total of 981 static facial expression images. Figure 5.2 shows the CK+ dataset samples. CK+ dataset is openly available for performing different experiments. Many researchers used this dataset for their research. In our methodology, we split the data by 90% and 10% in training and testing respectively. Total number of images in dataset present in each class are shown in table 4.2

Table 4.2: Distribution of the CK+ facial images

Expression	Sample
Angry	137
Disgust	177
Fear	72
Happy	213
Neutral	732
Sad	84
Surprise	242

4.1.3 Japanese Female Facial Expression (JAFFE)

On contrary, the Japanese Female Facial Expression (JAFFE) is a comparatively less dataset containing only 213 images of 10 Japanese female models and publicly available dataset for research purpose [96]. These images are used to express all those 7 facial expressions as that in FER2013. Some sample of JAFFE dataset are shown in Figure 4.3

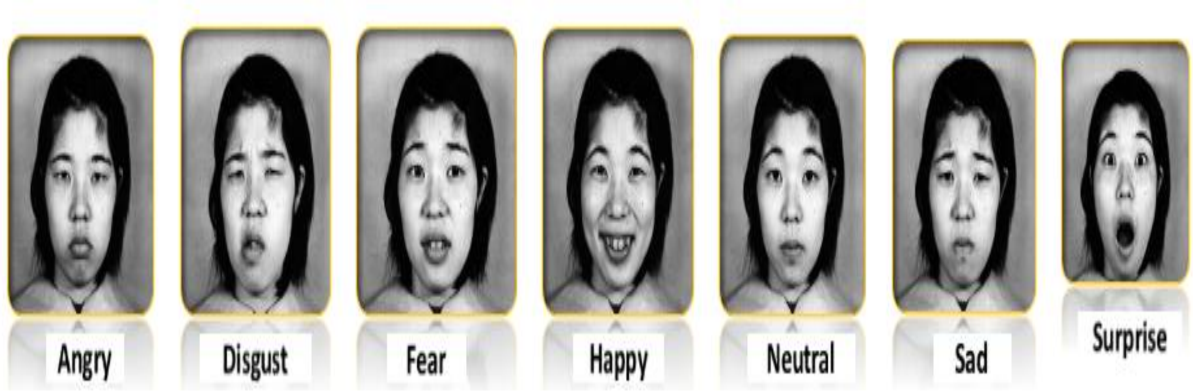


Figure 4.3: JAFFE Dataset Sample

The distribution of the facial images dataset in 7 classes are: 30 angry, 29 disgust, 33 fear, 30 happiness, 31 sad, 30 surprises and 30 neutral containing on average 3 or 4 images per subject. During the process of taking the images, the lighting was controlled strictly to ensure consistency and occlusion like a hair or eyeglass is not allowed. The original image has the following specifications: all expression in frontal view and 256 x 256 pixels. JAFFE is usually used under controlled lab environments; therefore, their application demands images to be front facial with limited variations. Static/still images are contained in the JAFFE.

4.2 Training Phase

Training phase include different steps including data pre-processing. Facial dataset is used to train CNN models. Data for training must be organized by their respective labels and prepared in such a way that they are consistent. This consistency approach guarantees that the network will only be trained on relevant extracted features without being distracted by things like background or objects in the background.

4.2.1 Image Pre-Processing

Pre-processing is one of the most important steps in any machine learning system, not just FER systems. As we know that any lack of screening can lead to unwanted, and wrongs results when raw data is analyzed. Therefore, quality assurance of data is essential before obtaining its relevant features. It is necessary to start our methodology with data preprocessing. Preprocessing of data is performed in order to eliminate noise and to normalize and centralize the gray scale of the image in order to ensure a firm foundation for future classification and identification.

Original dataset for FER2013 consists of an 8-bit grayscale vectorized .csv file with the corresponding labels. The dimension of a pixel vector is 48*48. There is one row for each picture. Initially the facial images are resized to 96*96 pixels and converted to the three-dimension image. After resizing the facial images, we have changed the image pixels into an array and data type float32. Finally, normalization is performed for faster convergence. The data are normalized into the range of [0, 1] this is achieved by normalizing each value by the input matrix of 256. Above mentioned all steps are performed on second CK+ and JAFFE dataset. Lastly The emotion-label is encoded by integers from zero to six (0 - anger, 1 - disgust, 2 - fear, 3 - happiness, 4 - sadness, 5 - surprise, 6 - neutral). Figure 4.4 shows seven classes with assign label. If categorical cross-entropy is used as a training metric, the label must be encoded in a K-of-one or a one-hot scheme. With the one-hot encoding, every label is represented by a single high level, 1, whereas the remainder of the classes are given by zeros. For example: For a label of class 3 (happy), the one-hot encoding is 0001000; for a label of class to 5 (sad), the one-hot encoding is 0000010, etc.

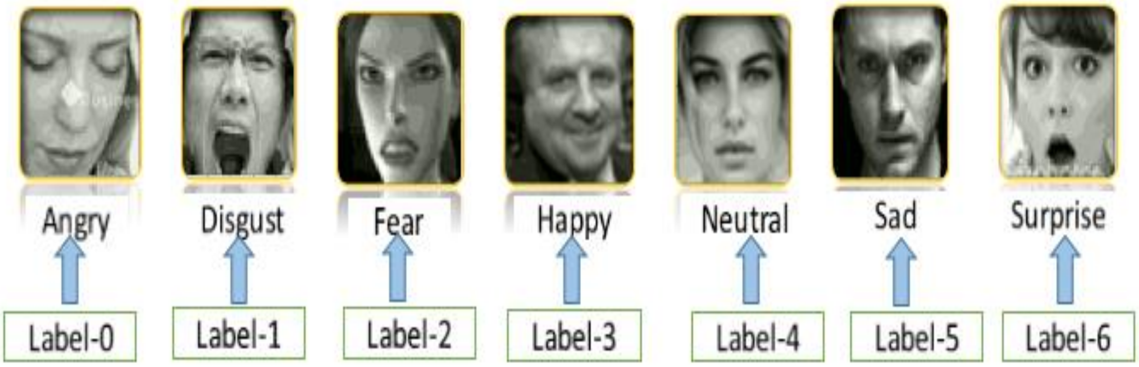


Figure 4.4: Seven Facial classes with assign labels

4.2.2 Data Augmentation

A common problem with emotion recognition databases is their small size, which makes them unsuitable for Machine Learning. Overfitting in Machine Learning models is often a problem when training on small data [97]. When the data used for training classified correctly it means model over fits, but in actual accuracy drops when it classifies data outside of the training set called bad generalization. Hence, during the model training overfitting can be detected easily: When the model is trained by using the training data, it gives very good accuracy but on validation data, the accuracy is very low. In order to solve this problem, which usually occurs in case of using small databases for training, data augmentation is used as a solution. Data augmentation is a technique in which we increase the dataset by modifying it in a way that is reasonable. These modifications can include smaller operations like cropping, rotating, flipping, brightness changing, zooming, rescaling, shifting and many others.

In our proposed methodology, we used three main properties of the data augmentation i.e., scaling, brightness and flipping. Data augment scaling property use to zoom in or zoom out the image. Each axis is scale independently. We have scaled the image below to 150% to 80% of the image height/width. In case of brightness, we adjust the image brightness using **GammaContrast** by scaling pixel values. Values in the range gamma= (0.5, 2.0) seem to be sensible. We have bright the image by gamma= (2.0) in our architecture. One of the most important data augmentation attributes is flipping. Image can be flip horizontally and vertically. Here, we have flipped the image by horizontally. Fliplr flip the image horizontally. Brightness property is applied to the scaling image and then flip that brighter image as describe in figure 4.5. Emotion labels are again assigned in this step as describe above.

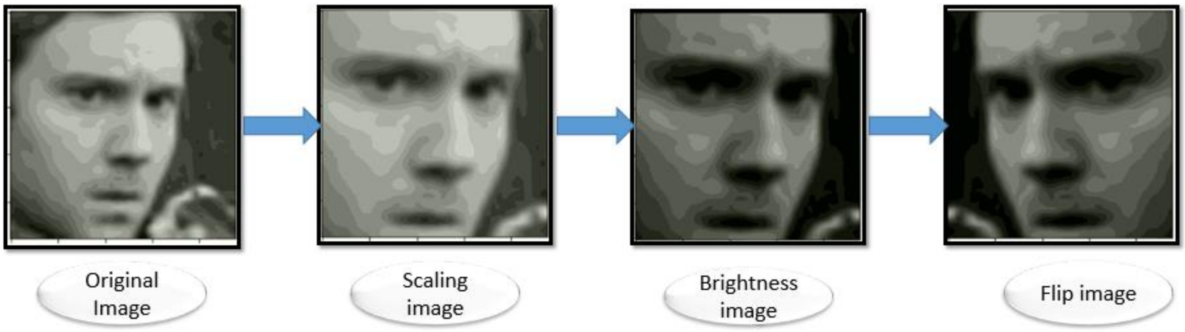


Figure 4.5: Facial image Augmentation sample of FER2013 dataset

Below table shows the augmentation data length of FER2013 dataset, CK+ dataset and JAFFE dataset.

Table 4.3: Augmentation on Datasets

Dataset	**Original Data**	**Augmentation Type**	**Augmented data**
FER2013	**32298**	Scaling->Brightness->Flip	32298
CK+	**1658**	Scaling	1658
		Brightness	1685
		Horizontal Flip	1658
		Total	**4974**
JAFFE	**213**	Scaling	213
		Brightness	213
		Horizontal Flip	213
		Total	**639**

4.1.3 Color Mapping

Methods such as color mapping or color transfer aim to recolor an image by establishing a mapping between an image and another image used as a reference [98]. The use of color as a means of conveying information or conveying a specific mood is an integral part of our visual world and one of the features of images we use in art, photography and visualization. Color mapping play an important role in visualization as they are able to improve the efficiency and effectiveness of data perception and therefore allow more insights into the data.

OpenCV defines 12 basic colormaps that can be applied to a grayscale image. So OpenCV use applyColorMap() function to produce pseudo colored image. Figure 4.6 shows the visual representation of colormaps. Lower grayscale values are replaced by colors to the left of the scale while higher grayscale values are on the right of the scale. We can apply different color maps to an image using the method applyColorMap().

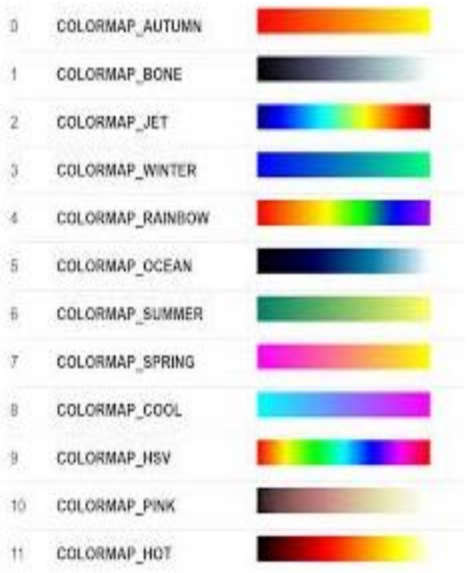


Figure 4.6: Visual representation of colormaps [99]

There are various other types of color maps which can be used for image visualization [99]. We have used various types of color maps and the best result is obtain from the COLORMAP_OCEAN. Figure 4.7 shows the pseudo-colored image sample.

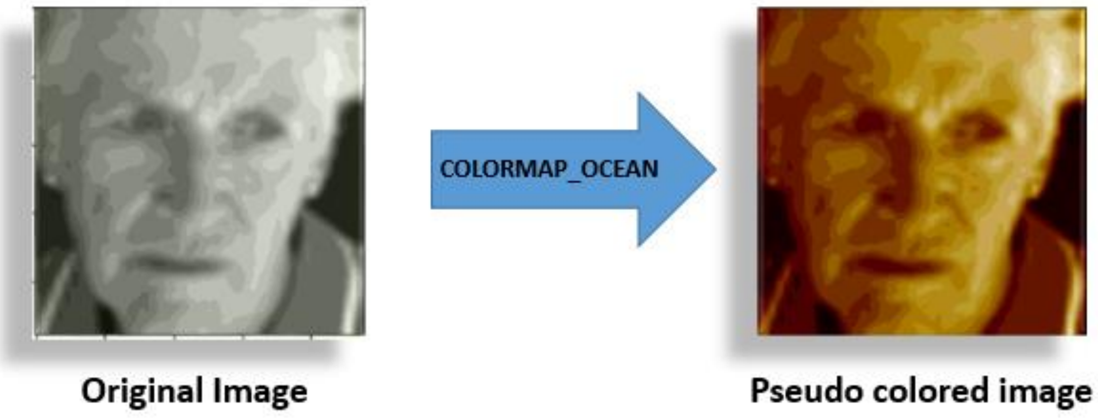


Figure 4.7: Pseudo colored image sample

4.2 Classification

We have described extensively in this section, how we use latest CNN-based deep learning architecture to categorize seven basic emotions of facial expressions. The FER database enlists the seven basic emotions (anger, fear, disgust, happiness, sadness, surprise, and a neutral expression). The FER databases are developed in controlled laboratory environments generally in a controlled lighting environment while others are built under uncontrolled environments. We have tested our proposed latest CNN-based deep learning architecture on uncontrolled database FER2013 which is taken from the 'Kaggle' competition "Challenges in Representation Learning: Facial Expression Recognition Challenge" and second dataset is controlled The Extended Cohn–Kanade database (CK+). Our proposed structure shows best result on both databases. The initial step of classification is training phase which is one of the most important phases.

4.2.1 VGG16

There are 16 convolution layers in VGG-16, and has a small receptive field is 3×3. In total there are 5 max pooling layers, each with a size of 2*2. After the last Max pooling layer, there are three fully connected layers. The final layer is the softmax classifier. ReLU activation is applied to all hidden layers. A schematic of the VGG-16 architecture is shown in Fig. 4.8

In convolution layers, training weights are automatically extracted and stored. As a result, the convolution layer and the FC store the weights of the training results. That allows them to determine the number of parameters to be used. The first to 19 layers are part of the extract-features layer, and the 20th to 23rd layers are part of the classification layer. In total, there are 138357544 original VGG16 parameters. A large dataset like ImageNet can be handled quite appropriately.

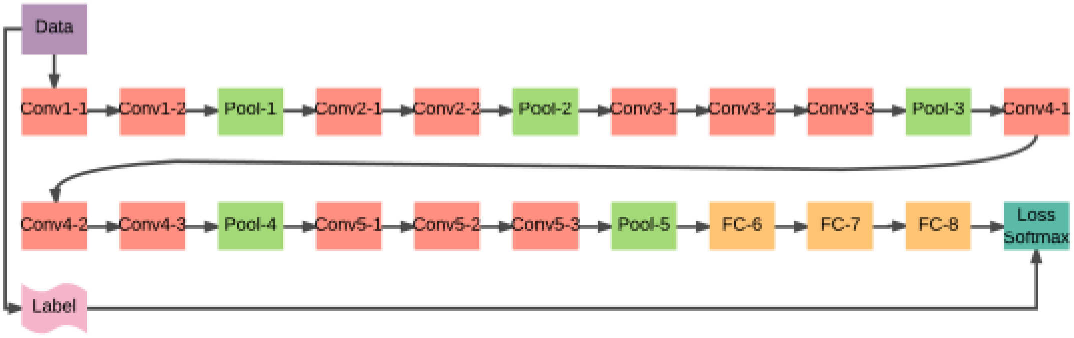


Figure 4.8: VGG16 Architecture [100]

4.2.2 DenseNet169

The DenseNet is one of the most effective convolutional neural networks, and it consists of very short connections between input and output layers. These networks have great potential to improve information flows and gradients throughout a network. The network allows us to produce reliable improvement in accuracy as the number of parameters increases, without degrading or overfitting the performance. DenseNet, therefore require significantly fewer parameters and reduce computation in order to achieve novel results [101]. The layered architecture of the DenseNet169 is shown in Figure 4.9

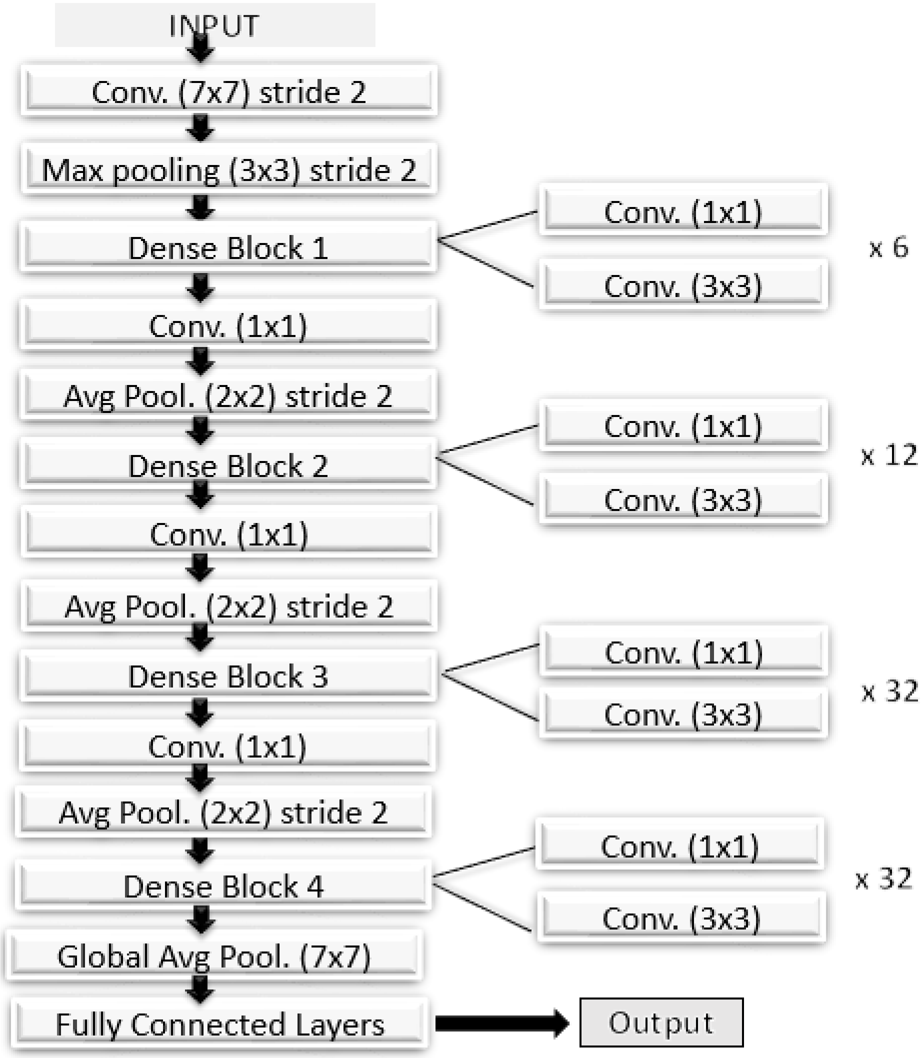


Figure 4.9: DenseNet169 Architecture

4.2.3 EfficientNetB0

Based on the study by Tan et al. [102] on the relationship between depth and width of CNN models, a way was found to design CNN models with fewer parameters that still achieve better classification accuracy. As a result, they named them EfficientNet CNN models and proposed seven different models in their original paper, labeled EfficientNetB0 to EfficientNetB7. Using the ImageNet dataset, Tan and Le [102] investigated EfficientNet CNN models and found that they outperformed all previous models both in terms of TOP-1 accuracy and parameter number. As compared to other models like ImageNet with same accuracy, EfficientNet is much smaller. For instance, the ResNet50 model has 23,534,592 total parameters in Keras application as shown, but yet it still underperforms the EfficientNet that is smallest also called EfficientNet-B0), which has 5,330,564 total parameters.

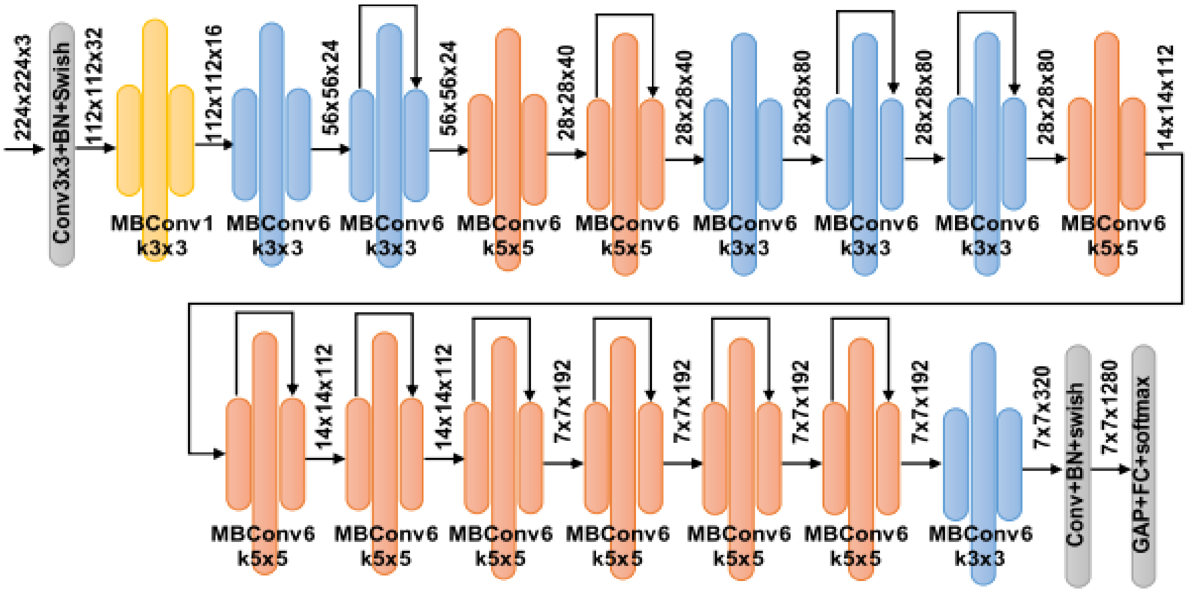


Figure 4.10: Basic EfficientNetB0 Architecture [103]

A new method for scaling up CNN models makes the EfficientNet family. The compound coefficient is simple and highly effective. In contrast to traditional methods that scale the depth, width, and resolution of networks, in EfficientNet each dimension is used to scale, a fixed set of scaling coefficients are used which are uniformly applied across all dimensions of the network. By balancing all dimensions of the network with respect to the available resources increases the performance of the network overall. By scaling the individual dimensions, we can improve the model performance, but scaling all dimensions simultaneously improves performance overall.

Furthermore, Swish activation function, rather than RELU, is used with this network as an activation function. In terms of shape, Swish activates in the same way as ReLU and LeakyReLU function and thus shares some of their advantages in terms of performance. However, it is a smoother activation function unlike these two.

$$f_{swish}(x) = \frac{x}{1+e^{-\beta x}} \quad \ldots\ldots\ldots\ldots\ldots\ldots.. (4.1)$$

During the training of the CNN model, β ≥ 0 is a parameter that can be learned. Note, if β = 0, f_{swish} is taken as the linear activation function and as β → ∞, f_{swish} looks more like the ReLU function except it is smoother.

From figure 4.10 there are some observations. The first observation is that MBConv1, MBConv3, and MBConv6 blocks are repeated in this bassline model[104]. The MBConv blocks basically come in different types. Second, we observe that the number of channels within each block increases or expand (through the use of more filters). Thirdly, inverted residual connections are being observed between the narrow layers of the model.

Mobile inverted bottleneck convolution (MBConv) block

The main building block of EfficientNet model family is the mobile inverted bottleneck convolution (MBConv). MBConv borrows concepts from the MobileNet models [52].

A main idea is to separate convolution layers based on depth, which consists of separate depth and point convolution layers. Then two more ideas are borrowed from MobileNet-V2 (which is a second improved version of MobileNet) including 1) inverted residual connections, and 2) Linear bottlenecks. According to the earlier discussion, the model scaling idea depends strongly on a network's baseline. The MBConv6 is shown in figure 4.11.

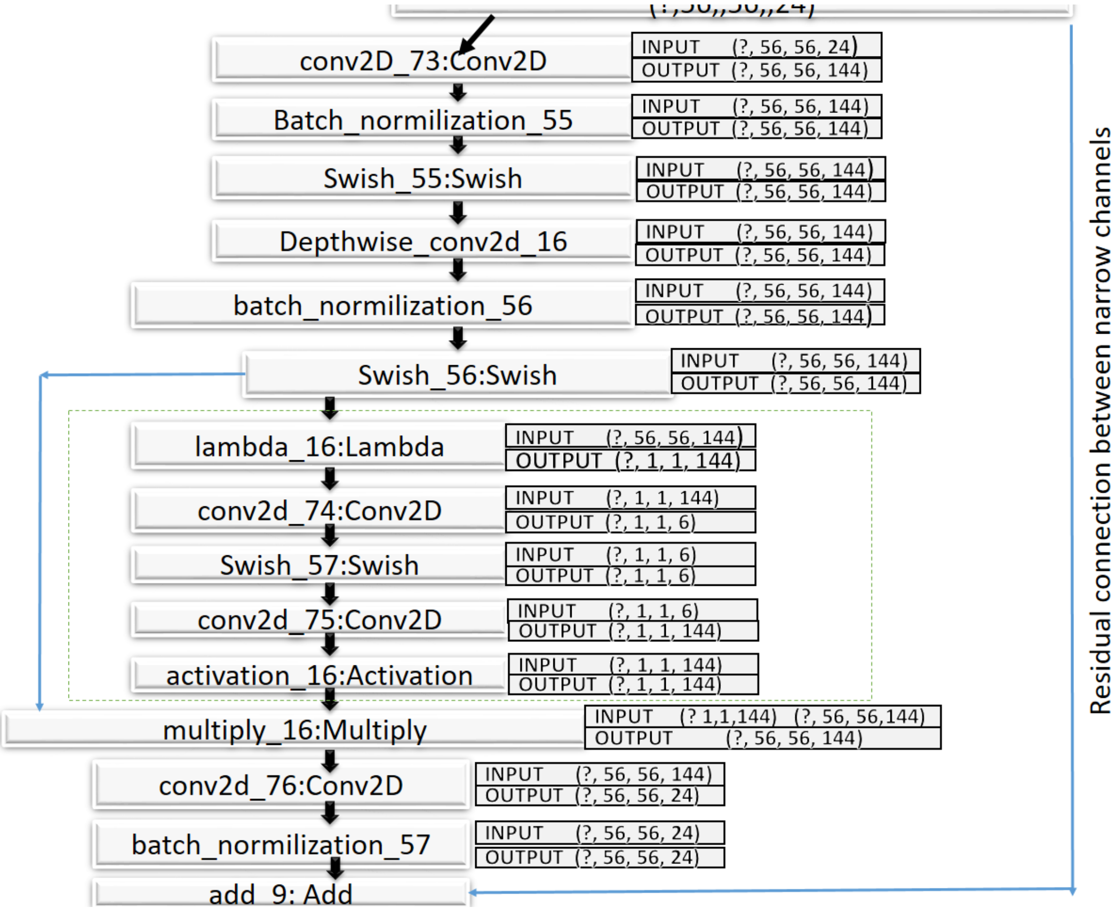


Figure 4.11: MBConv6 Architecture [103]

Another type, shown in Fig. 4.12, is called MBConv1 which is used at the beginning of the EffeicnetNet models. Furthermore, each type can be further divided into several variants depending on the filter size of the convolutional layers (which could be 3*3 or 5*5) and whether the block contains an inverted residual connection.

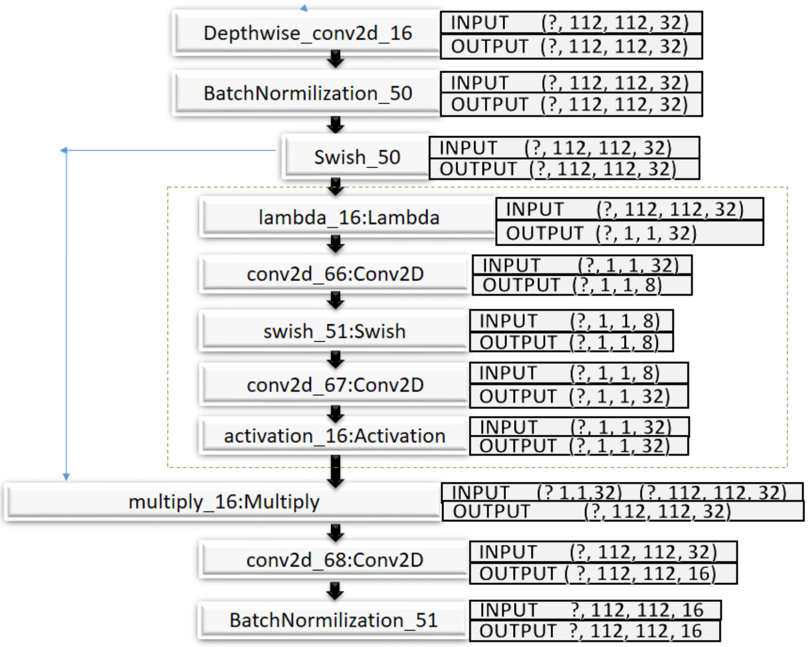


Figure 4.12: MBConv1 Architecture [103]

4.2.2 Additional Layers in EfficientNetB0

We explore the EfficientNet by adding more operator blocks on below of it. More specifically, we add four new blocks, as shown in figure 4.8. For us to learn from our prepared data, we need to use the pre-trained EfficientNetB0, in our system we have proposed to replace the EfficientNetB0 baseline model's the final layers and add our set of layers to activate achieve weights from it. Figure 4.5 illustrates EfficientNetB0 composed of a Global Average Pool (GAP) with our proposed final layers for it, a sigmoid classifier and three FC dense layers. We have added a pooling layer to prevent overfitting from the sophisticated feature handling. The rescaling of the height, width, and depth of the incoming tensor from the base model is done by using the GAP layer which further reduces the number of parameters into a 1x1x3 dimension, respectively. Before predicting results, we directed the feature sets from the previous GAP to a set of dense layers consisting of 512 hidden units which was connected to another dense layer with 256 hidden units.

Also, a Rectified Linear Unit (ReLU) is applied for the activation of the hidden layer. At the end Sigmoid activation function is used in which a binary class classification is specifically performed by using a logistic function [104]. The S-shaped non-linear function binds values to a 0 or 1.

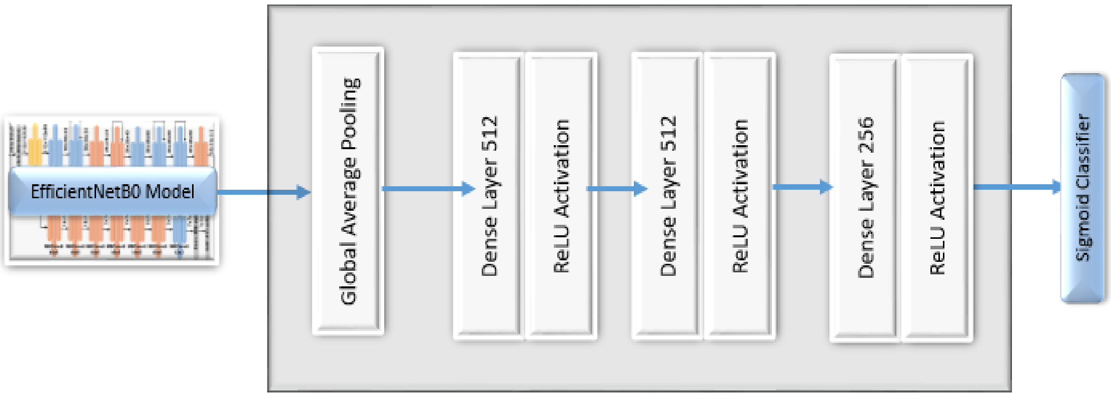


Figure 4.13: Proposed Final Model

4.2.3 Training and Inference

We use an EfiicientNetb0 baseline model with some additional layers network. The original images are scaled to 96*96 and biased to the range of [0, 1]. For FER2013 database, we follow the data augmentation technique for training: the image is zoom out with scaling property, and augmented scale image is brighter by p=2 and final flip the image horizontally. In FER2013 database, data are already divided into training, validation and testing sample. The model is trained with different batch size, the mini-batch size is 16 and totally 100 epochs. Learning rate η is varies from 0.1 to 0.00000001. The best rest achieve with learning rate of 0.000001 and batch size 50.

Table 4.4: Best Training Parameter

Parameter for Training	**Value**
Optimizer	Adam
Batch Size	5

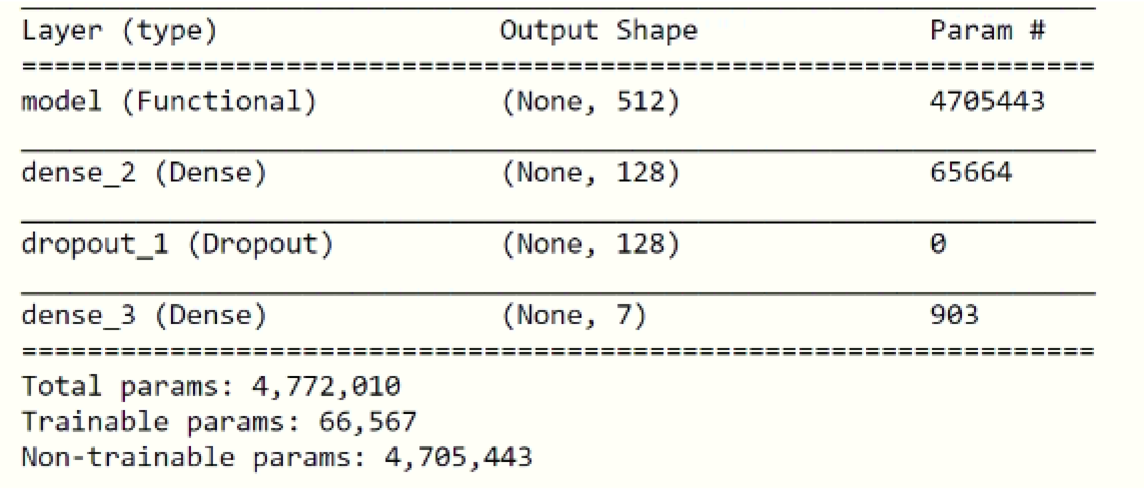
Learning Rate	0.000001
EPOCHS	50, 100, 200

We use Adam optimizer and sigmoid activation function for training purpose. As for CK+ database, we use different data augmentation technique while the remaining process are same as discussed above. CK+ is a small database as compared to the FER2013, so we have applied three data augmentation properties separately and then concatenate all augmented facial images. In CK+ database, we spit the data in 80% and 20% for training and testing purpose respectively.

4.3 Modification of EfficientNetB0 with additional layer Architecture

When evaluated using the current databases, FER models (including those using deep neural networks) don’t achieve desire accuracy. Moreover, most related works (especially those that use cross-database protocol) don’t carry out an extensive experiment (usually training in a single database and testing in another one, i.e., using few databases). Consequently, these methods are also difficult to generalize.

In our proposed methodology, we froze the beginning layers on the base model, then trained our proposed model with some modification in ending layers with the FER2013 database. Since we have dealt with an unbalanced data set, the proposed architecture may have a huge possibility to confront the problem of overfitting. Some modification in our proposed can avoid overfitting and form a generalized model. We have experimented with different modification layers techniques to get maximum accuracy results. Finally, we have finalized the one technique in which we have to freeze the upper layers and weights of the pre-trained model until FC1 and, after that we unfreeze the layers and replaced them with the new output layer and activation map. After applying modification mechanism on above define model, total number of parameters are 4,772,010 out of which only trainable parameter are 66,567 and non-trainable parameters are 4,705443. The modify model summary is shown in figure 4.14



```
Layer (type)                 Output Shape              Param #
=================================================================
model (Functional)           (None, 512)               4705443

dense_2 (Dense)              (None, 128)               65664

dropout_1 (Dropout)          (None, 128)               0

dense_3 (Dense)              (None, 7)                 903
=================================================================
Total params: 4,772,010
Trainable params: 66,567
Non-trainable params: 4,705,443
```

Figure 4.14: Modify model Summary

4.3.1 Training and Inference for modify model

Above describe modify model is train on augmented facial image with color mapping and original augmented facial images. Total number of facial images on which modify model train are 64596. We have trained our model with variable learning rate, EPOCH and batch size. But the best results are achieved with the same parameter which are used for the training of the above-mentioned model. All the data augmentation techniques are only applied on the training data.

4.4 Testing Phase

For testing purpose, it is also necessary to start with data preprocessing, the general purpose of data preprocessing is to eliminate noise and to normalize and centralize the gray value of the image to provide a solid foundation for subsequent testing. As mentioned above, the test facial images are also resized to 96*96. Once image is resized, the facial image pixels are changed into an array and data type float32 and at the end data is normalized in the range of [0, 1]. Emotion labels are assigned as per figure 4.1 and covert to one-hot encoding scheme. Data augmentation is not applied on the test images data.

4.5 Proposed Architecture

In our proposed architecture, we performed image pre-processing on FER2013 facial images, and that pre-processed image is augmented with scaling augmentation and the brightness feature is applied on it and finally brighten facial image is horizontally flip. Color mapping is applied on the pre-processed facial images. Augmented data and original pre-processed data concatenate in a signal array of image and lastly, we concatenate color mapping image with previous mentioned array of images. At this stage total number of training facial image in array are 96,894. Then we train the model on training facial image array along with assigned labels with learning rate 0.0000001, batch size 50 and epoch 100. Figure 4.15 shows the complete architecture of our proposed model.

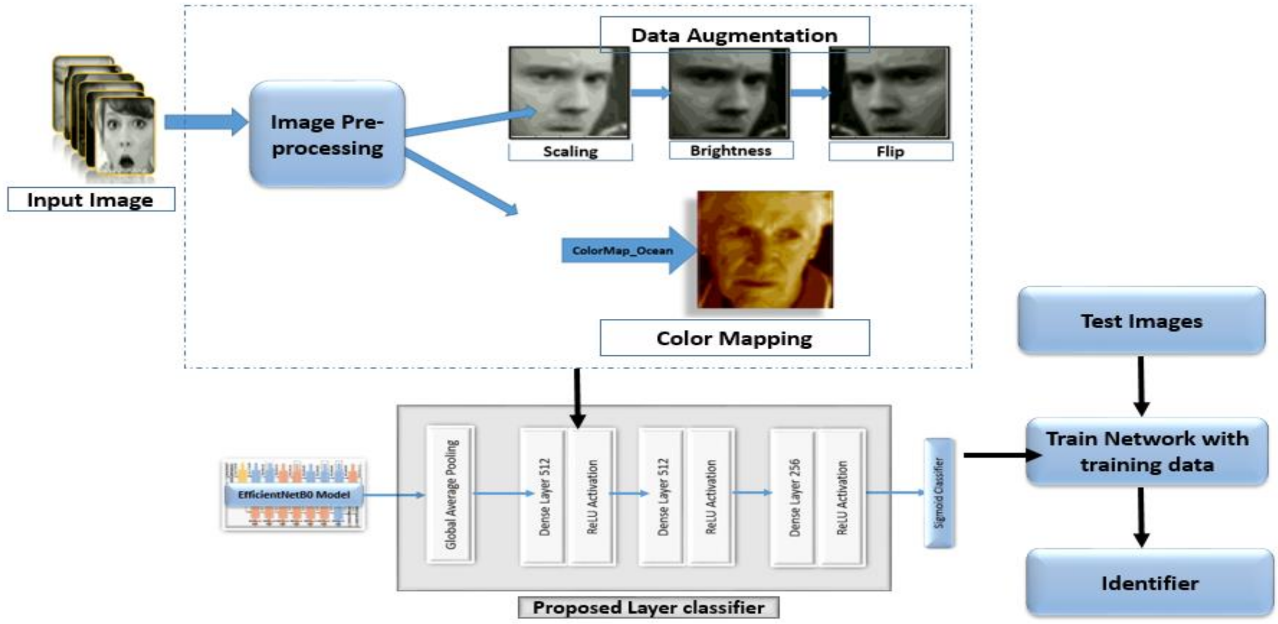
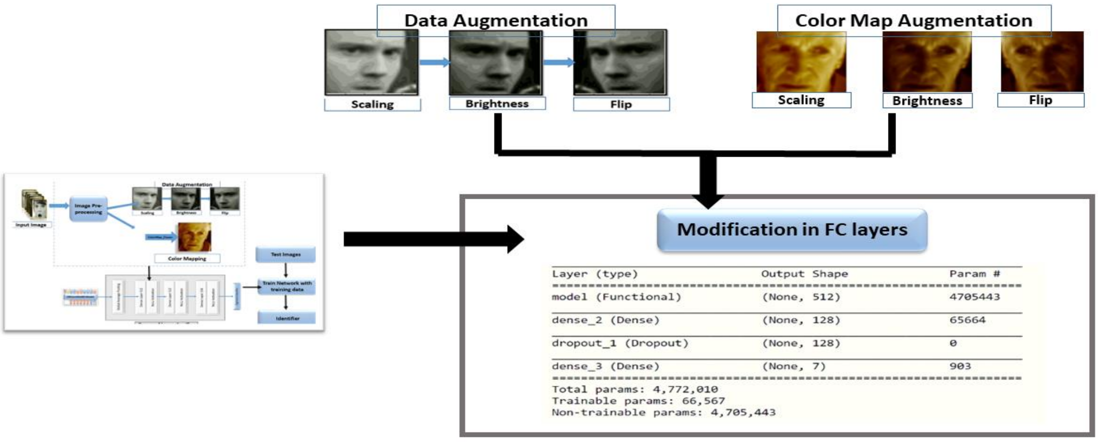


Figure 4.15: Flow diagram of proposed Methodology

The final and most important step of our model is modification in FC layers in which the upper layers and weights of the pre-trained models are frozen until fully connected layer-1 and then we add new output layers in the model. The modify model is train on the same parameters.

Augmentation data and color mapping augmentation data is concatenate in single array and use to train the model as shown in figure 4.16.



```
Layer (type)                 Output Shape              Param #
=================================================================
model (Functional)           (None, 512)               4705443
_________________________________________________________________
dense_2 (Dense)              (None, 128)               65664
_________________________________________________________________
dropout_1 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_3 (Dense)              (None, 7)                 903
=================================================================
Total params: 4,772,010
Trainable params: 66,567
Non-trainable params: 4,705,443
```

Figure 4.16: Modification in FC layer

Chapter#5

Experimental Results

CHAPTER 5: EXPERIMENTAL RESULTS

In this chapter, the results are discussed according to the pre-processed dataset. Our analysis used a confusion matrix to identify the misclassified samples and the number of correctly classified samples. Based on this analysis we computed our proposed model accuracy.

5.1 Results

The previous section described in detail the databases. Now we will analyze how the results of the experiment were evaluated. In this section, the ideas of our design are evaluated on three different datasets. Hence, different feature representation methods and backbones are tested, and the results are discussed.

With the following method, we compute facial recognition accuracy. The accuracy is calculated via

$$Accuracy = \frac{(TP + TN)}{TP + TN + FP + N} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots . (5.1)$$

Where true positive (TP), true negative (TN), false positive (FP) and false negative (FN).

5.1.1 FER2013 Dataset Results

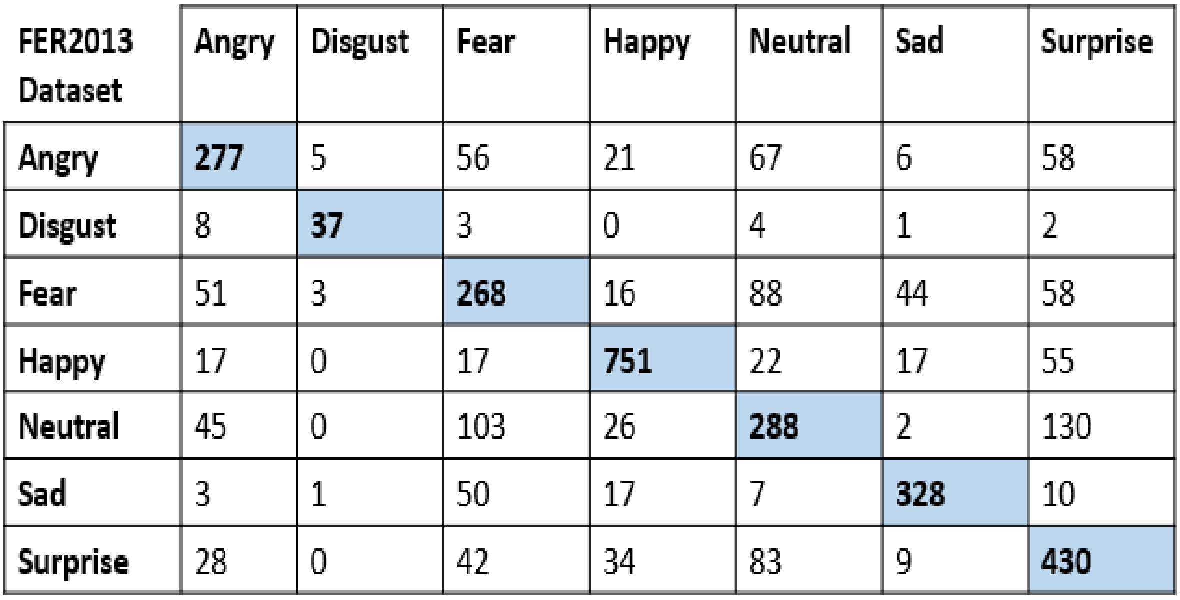
The model proposed by using EfficientNet-B0 is evaluated on the FER2013 data. Using network parameters, the maximum Epoch size is 50 for high-resolution images with learning rate 0.0000001. Series of experiment are applied on FER2013 dataset using different network parameter and measure the facial recognition accuracy in order to choose best optimizer, loss function and coloring mapping property. Table 6.1 demonstrate the comparative analysis of proposed EfficientNetB0 model with different CNN model.

Table 5.1: Comparative analysis of proposed efficientNetB0 with different CNN model

CNN- Deep Learning Model	Testing Accuracy
Simple EfficientNetB0	82.2%
EfficientNetB0 + VGG16	90.4%
DenseNet169 with additional layers	89.90%
DenseNet169 with modify layers	91.60%
EfficientNetB0 with additional layers	90.87%
EfficientNetB0 with modify layers (Proposed Architecture)	**94.88%**

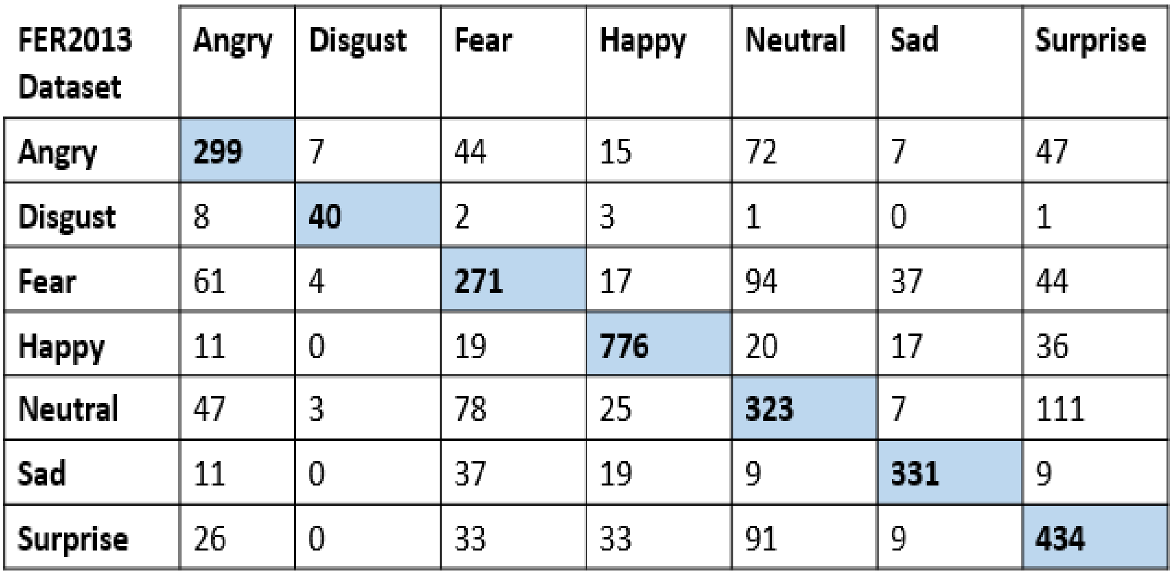
Specifically, results are shown in Table 5.1 where FER2013 datasets are trained with the data enhancement operation named augmentation by the Combine model of EfficientNet and VGG-16, DenseNet169, DenseNet169 with some modification in last layers, EfficientNetB0 and finally EfficientNetB0 with modify model. Comparing mentioned model accuracy, it can be seen that the EfficientNetB0 with modify model is better with testing accuracy of 94.88% than the remaining classifier in case of FER2013 dataset. The results are significantly improved with the modification in last layers technique in both CNN-deep learning architecture of efficientNetB0 and DenseNet169.

It is observed that the accuracy for all the model is decreasing in the order of EfficientNetB0 + modification in last layers technique (94.88% for FER2013) > DenseNet169 + modification technique (91.60% for FER2013) > EfficientNetB0 (90.87% for FER2013) > EfficientNetB0 + VGG16 (90.4% FOR fer2013), indicating that the EfficientNetB0 + modification in last layers technique is better DCCN model for FER2013.



FER2013 Dataset	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	**277**	5	56	21	67	6	58
Disgust	8	**37**	3	0	4	1	2
Fear	51	3	**268**	16	88	44	58
Happy	17	0	17	**751**	22	17	55
Neutral	45	0	103	26	**288**	2	130
Sad	3	1	50	17	7	**328**	10
Surprise	28	0	42	34	83	9	**430**

Figure 5.1: Confusion matrix using proposed EfficientNetB0 model with FER2013 dataset



FER2013 Dataset	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	**299**	7	44	15	72	7	47
Disgust	8	**40**	2	3	1	0	1
Fear	61	4	**271**	17	94	37	44
Happy	11	0	19	**776**	20	17	36
Neutral	47	3	78	25	**323**	7	111
Sad	11	0	37	19	9	**331**	9
Surprise	26	0	33	33	91	9	**434**

Figure 5.2: Confusion matrix using proposed EfficientNetB0 + Modify model with FER2013 dataset

The confusion matrix of FER2013 dataset with EfficientNetB0 and EfficientNetB0 with modification are shown in figure 5.1 and 5.2 respectively. The total number of correctly identify angry class sample are 299 and the overall accuracy for angry class is 90%. In case of disgust class, the accuracy is 96% with correctly identify sample 40 out of 55. The properly classify facial image for fear and happy class are 271 and 776 respectively with accuracies 86% and 97%. The most misclassify label in this dataset is fear class. Moreover, for neutral, sad and surprise the accuracies are 87%, 96% and 90% respectively with 323, 331 and 434 correct identify sample of each class. It can be seen more clearly in ROC graph shown in figure 5.6

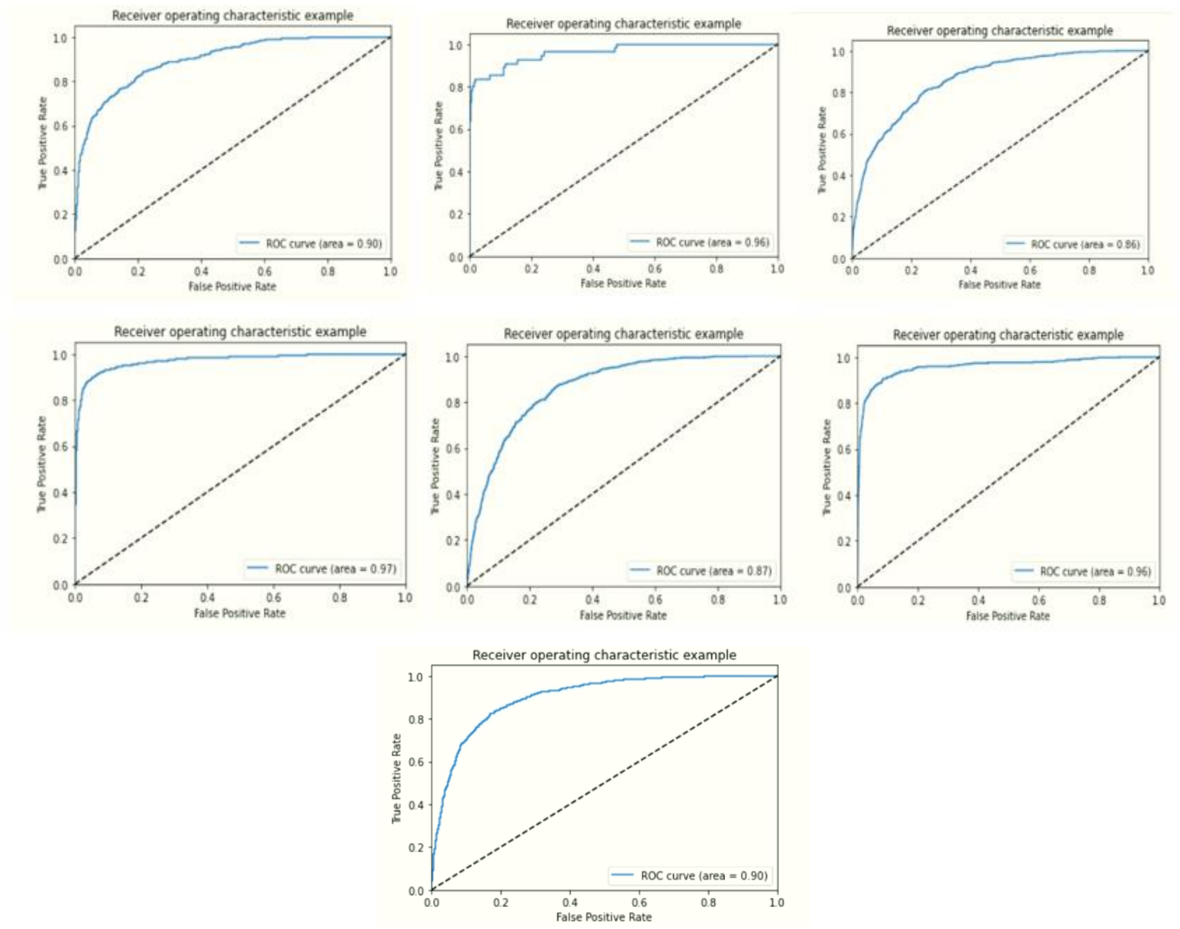


Figure 5.3: Individual class accuracy using proposed EfficientNetB0 + modify Model with FER2013 dataset

Comparison of FER2013 dataset with state-of-art

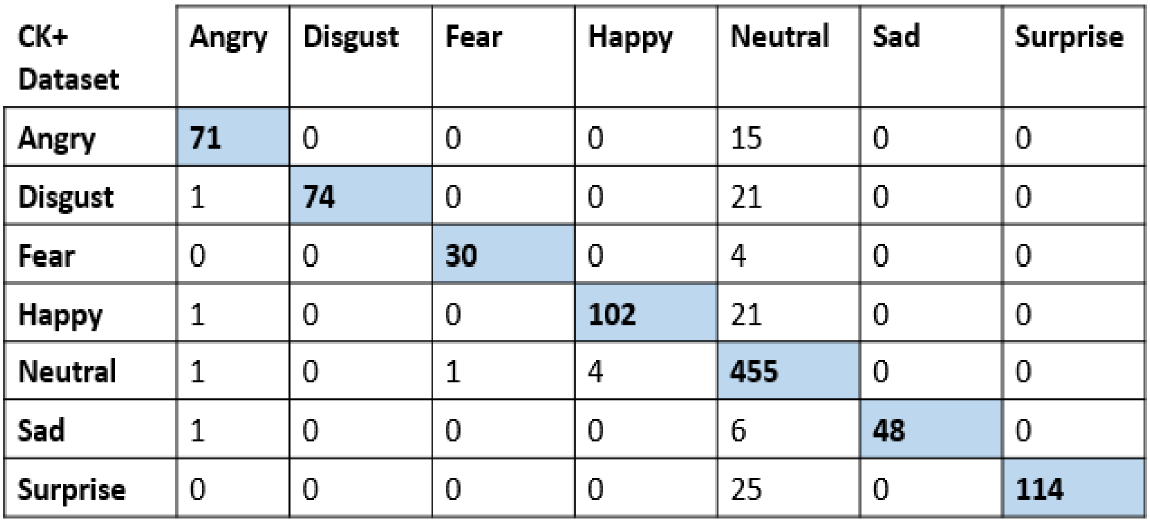
Table 6.2 compares the results of a recent study performed on FER2013 dataset. Yichuan Tang [77] is the winner of the ICML competitions conducted in 2013 worked with Linear Support Vector Machines model and achieve 71.02% accuracy. In 2021 Khaireddin et al [71] has used VGGNet model with different optimizer and achieve 73.06% accuracy. Shiqian Li et al [74] worked on light-weight FER Network with Auto-FERNet along with effective relabeling method and the recognition rate that they achieved was 73.78%. Jun Liu et al [73] worked on new deep learning model that combine VGG and ResNet and obtained 94.5% facial accuracy which is the highest accuracy achieve on this dataset till now. Our proposed model achieved 94.88% accuracy.

Table 6.2: Compares the results of a recent study performed on FER2013 dataset.

Reference	Accuracy
Yichuan Tang (FER2013 Winner) [77]	71.20%
Prateek Chhikara et al [80]	71.64%
Khaireddin et al [71]	73.06%
Shiqian Li et al [74]	73.78%
Nguyen et al [83]	74.09%
Soodamani Ramalingam et al [86]	78%
Zhi Li [89]	85.20%
Huanhuan Ran, Shiping Wen et al [78]	89.92%
Jun Liu et al [73]	94.5%
Our proposed model	**94.88%**

5.1.2 CK+ Dataset Results

The second dataset which we used to evaluate the performance matrix of our proposed model is CK+ dataset. In FER2013, we achieve best results with efficientNetB0 with modify model so CK+ dataset is train only on that model. We have trained our model with network parameter like learning rate 0.001 and batch size 20 along 50 epochs. The overall facial expression recognition accuracy using our proposed architecture on CK+ is **99.16%.** Figure 5.4 shows the confusion matrix for CK+ dataset.



CK+ Dataset	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	**71**	0	0	0	15	0	0
Disgust	1	**74**	0	0	21	0	0
Fear	0	0	**30**	0	4	0	0
Happy	1	0	0	**102**	21	0	0
Neutral	1	0	1	4	**455**	0	0
Sad	1	0	0	0	6	**48**	0
Surprise	0	0	0	0	25	0	**114**

Figure 5.4: Confusion matrix using proposed EfficientNetB0 + modify model with CK+ dataset

From the confusion matrix, it is concluded that the most sample are misclassified in neutral class as already mentioned in section 4.1.2 that each display began in a neutral face expression.

Total number of test sample in angry class is 86 out of which 71 are classify properly and remaining misclassify as neutral expression. In case of disgust class, 74 sample are correctly identify out of 96 with 98% individual class accuracy. 99% sample are properly classified in fear and happy class. In neutral expression, 455 facial images are correctly identified out of 461, only 6 samples are misclassified in this case. For sad and surprise class, 48 and 114 sample are properly identify out of 55 and 139 respectively. Individual class accuracy using proposed EfficientNetB0 with modify model on CK+ dataset is shown in figure 5.5.

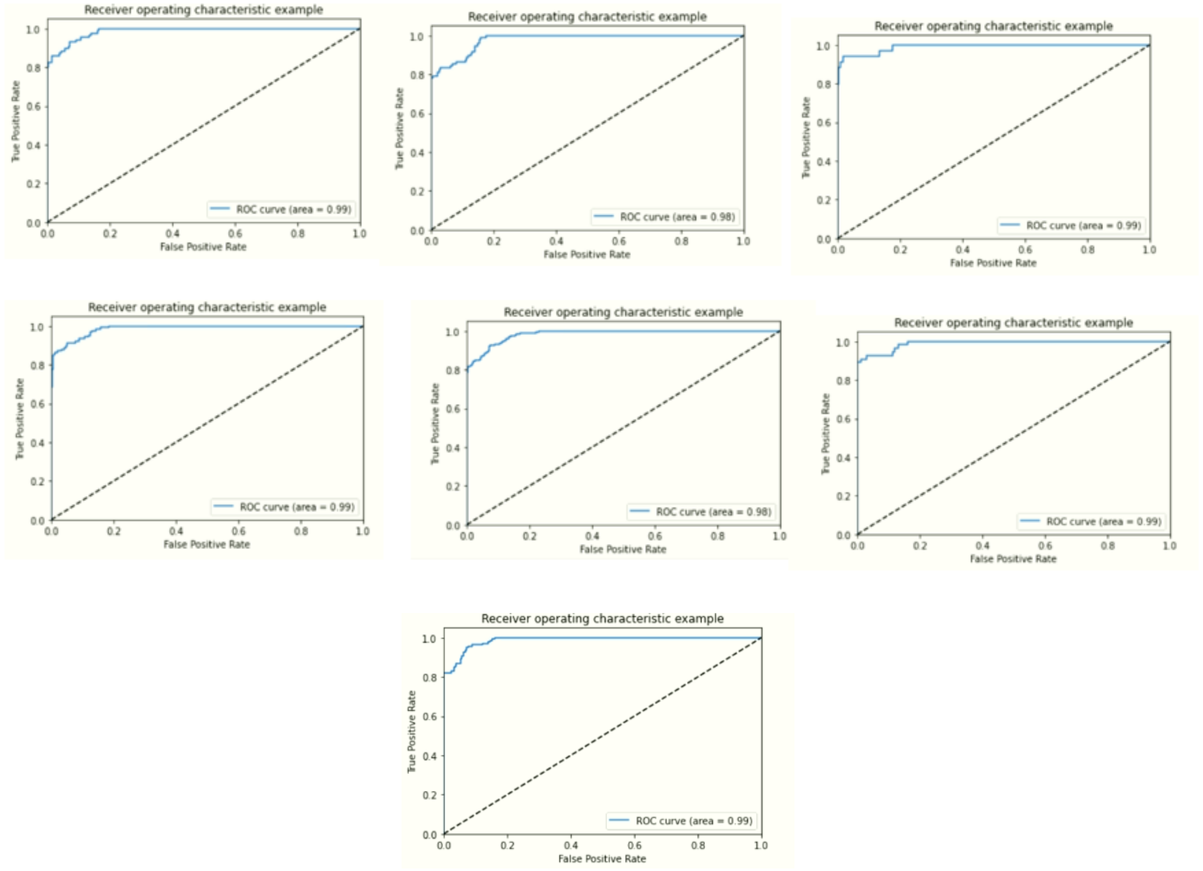


Figure 5.5: Individual class accuracy using proposed EfficientNetB0 + modify model with CK+ dataset

Comparison of CK+ dataset with state-of-art

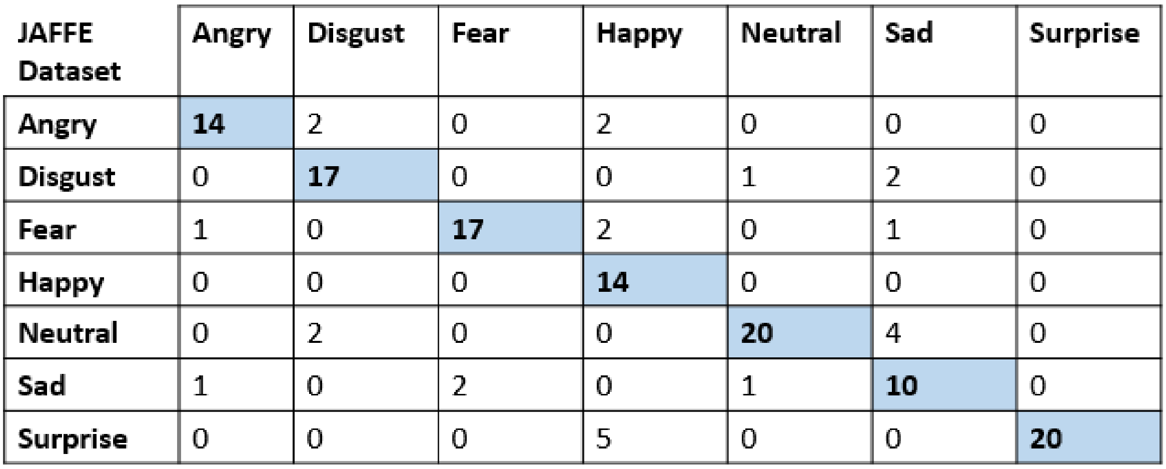
The performance of our model is compared against the state-of-the-art methods as shown in Table 6.2. Our method shows remarkable results. Mouath Aouayeb et al [105] introduced learning vision transformer with squeeze and excitation for facial expression and achieves 99.8% accuracy with CK+ dataset. Frame attention network model proposed by Debin Meng et al [106] and achieves 99.7% accuracy. Hui Ding et al [107] worked on regularizing a deep face recognition net for expression recognition. Yuedong Chen et al [108] proposed a novel FER framework, named Facial Motion Prior Networks (FMPN). Table 6.2 compares CK+ state-of-art performance.

Table 6.2: Compares the results of a recent study performed on CK+ dataset.

Reference	Accuracy
Mouath Aouayeb et al [105]	99.8%
Debin Meng et al [106]	99.7%
Hui Ding et al [107]	98.6%
Yuedong Chen et al [108]	98.06%
Shervin Minaee et al [109]	98%
Philip Lu et al [90]	95.70%
Ali Mollahosseini et al [91]	93.2%
Our proposed model	**99.16%**

5.1.3 JAFFE Dataset Results

Moreover, we have trained our proposed model on JAFFE dataset and evaluate the performance. Like FER2013 and CK+, we achieve best results with efficientNetB0 + modify model on JAFFE dataset also. We have trained our model with network parameter like learning rate 0.001 and batch size 20 along 50 epochs. The overall facial expression recognition accuracy using our proposed architecture on JAFFE is **98.97%.** Figure 5.6 shows the confusion matrix for JAFFE dataset.



JAFFE Dataset	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	**14**	2	0	2	0	0	0
Disgust	0	**17**	0	0	1	2	0
Fear	1	0	**17**	2	0	1	0
Happy	0	0	0	**14**	0	0	0
Neutral	0	2	0	0	**20**	4	0
Sad	1	0	2	0	1	**10**	0
Surprise	0	0	0	5	0	0	**20**

Figure 5.6: Confusion matrix using proposed EfficientNetB0 + modify model with JAFFE dataset

Figure 5.6 shows the confusion matrix of the JAFFE dataset. In JAFFE dataset, mostly facial expression is misclassified as a happy expression. In angry class, 14 samples are correctly identified out of 18 with 99% accuracy. In disgust and fear class, the properly labeled sample are 17 for each class out of 20. All happy class expression is correctly identified with 100% class accuracy. Moreover, neutral, sad and surprise class properly classified sample are 20, 10 and 20 out of 24, 14 and 25 respectively. The below figure shows the individual class accuracy graph.

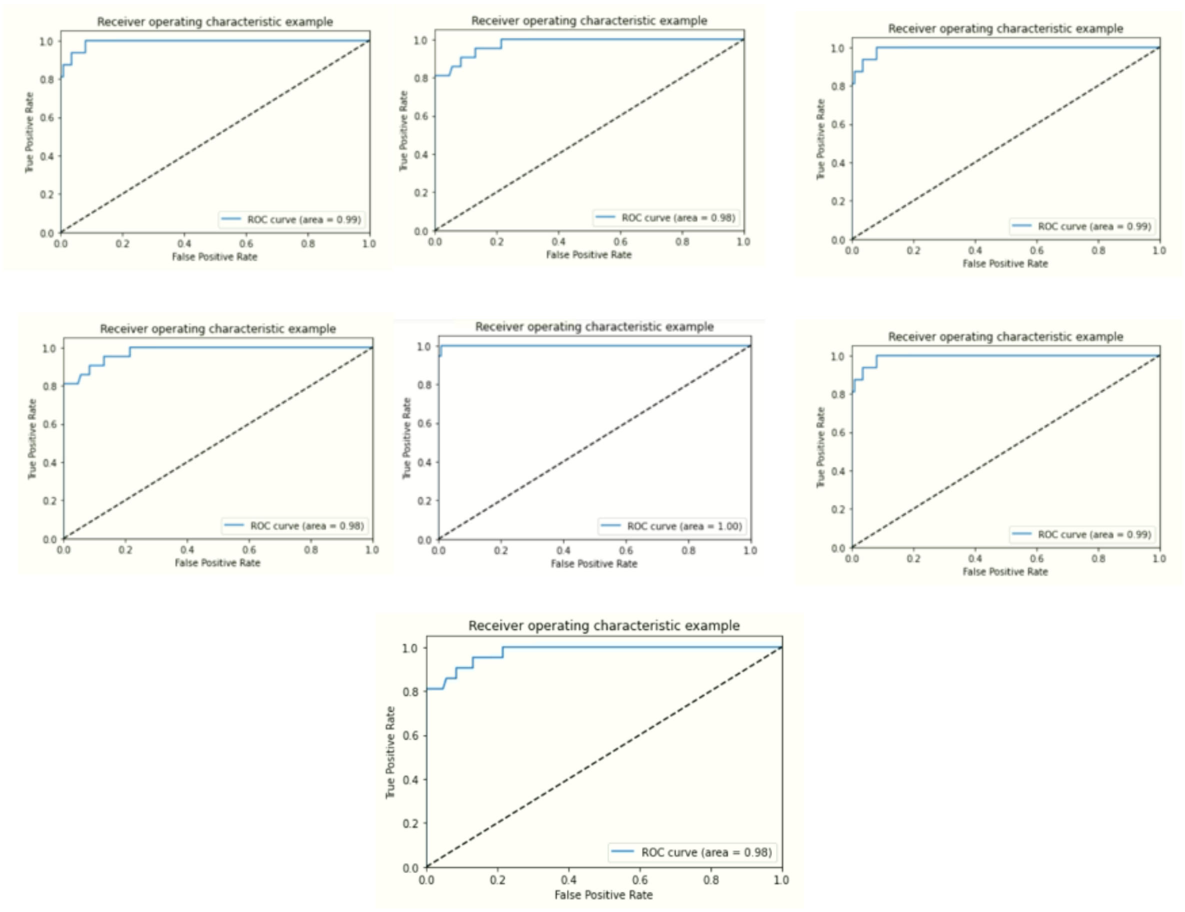


Figure 5.7: Individual class accuracy using proposed EfficientNetB0 + modify model with JAFFE dataset

Comparison of JAFFE dataset with state-of-art

S L Happy et al [110] proposed a new framework for expression recognition by using appearance features of selected facial patches. Yoshihiro Shima et al [112] proposed image augmentation of facial expression recognition on the basis of a mixture of a deep neural network and a support vector machine. Chang Liu et al [113] proposed a network that involves of a Spatial Attention Convolutional Neural Network (SACNN) and a sequence of Long Short-term Memory networks with Attention mechanism (ALSTMs) and achieve 98.57% accuracy with JAFFE dataset. Wenming Zheng et al [114] used Kernel Canonical Correlation Analysis (KCCA) for facial expression recognition and achieve 98.36% accuracy with JAFFE dataset. Jenni Kommineni et al [115] proposed a hybrid model sing dual-tree m-band wavelet transform (DTMBWT) algorithm based on energy, entropy, and gray-level co-occurrence matrix (GLCM).

Table 6.3: Compares the results of a recent study performed on JAFFE dataset.

Reference	Accuracy
S L Happy et al [110]	91.8%
Shervin Minaee et al [111]	92.8%
Yoshihiro Shima et al [112]	95.31%
Chang Liu et al [113]	98.57%
Wenming Zheng et al [114]	98.36%
Jenni Kommineni et al [115]	99.53%
Our proposed model	**98.97%**

Chapter#6

Conclusion and Future work

CHAPTER 6: Conclusion and Future work

6.1 Conclusion

Various methods and algorithms have been investigated for improving the recognition accuracy of facial expression recognition from images. Several methods have been proposed to detect the facial expressions in the images. Most of the methods use the laboratory-controlled facial expression datasets, which have controlled conditions. The lighting is uniform, and the images have an entire frontal face. The laboratory-controlled images have no occlusions in most of the cases. So, face detection and feature extraction process get easier as a part of the facial expression recognition process. Therefore, facial expression recognition on such datasets becomes a lot simpler than for real-time facial expression datasets. For the latter type of datasets, the images are taken from the internet and real-world images. Therefore, they have problems like a difference in lighting conditions, varying head poses, resolutions of images, and various occlusions like sunglasses, hairs etc.

The thesis's overall goal is to develop efficient models for facial expression recognition using deep learning techniques to achieve better recognition accuracy on lower-resolution images of real-time facial expression dataset for recognizing seven basic facial expressions such as happy, disgust, surprise, anger, sad, fear and neutral. In this research work, we have proposed deep learning model for recognizing facial expressions from images. To improve recognition accuracy of lower resolution images for real-time facial expression dataset and for the laboratory-trained facial expression datasets.

Hence our contribution in this research is the investigation and development of novel proposed models for improving the performance of facial expression recognition task using deep learning techniques on real-time facial expression datasets as well as on laboratory-trained facial expression datasets.

6.2 Future Work

Among all the works presented here in this thesis, there are areas to progress and improve further. Putting aside what we have successfully achieved, several useful extensions that can be addressed to further improvements as explained below:

1. Without considering the influence of head pose variations, only frontal faces are taken for training and implementation purpose. So, further faces from several views can be considered from the images or videos which may help to improve the recognition accuracy.
2. We have considered Appearance-based features for our research work. So, a hybrid method can be developed in the future by combining geometric features and appearance-based features to improve the performance of the facial expression recognition system.

3. REFERENCES

[1] Shao, J., & Qian, Y. (2019). Three convolutional neural network models for facial expression recognition in the wild. Neurocomputing, 355, pp. 82-92

[2] C. Ding and D. Tao, ‘‘A comprehensive survey on pose-invariant face recognition,’’ ACM Trans. Intell. Syst. Technol., vol. 7, no. 3, pp. 1–42, Apr. 2016.

[3] F. D. Guillen-Gamez, I. Garcia-Magarino, J. Bravo-Agapito, R. Lacuesta, and J. Lloret, ‘‘A proposal to improve the authentication process in mhealth environments,’’ IEEE Access, vol. 5, pp. 22530–22544, 2017.

[4] Ekman, P., & Keltner, D. (1997). Universal facial expressions of emotion. Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture, pp. 27-46

[5] Biometrics and Experimental Consumer Psychology"https://www.endlessgain.com/blog/biometrics-and-experimental-consumer-psychology-part-5-facial-expression-analysis/"

[6] Fathima, A., & Vaidehi, K. (2020). Review on facial expression recognition system using machine learning techniques. In Advances in Decision Sciences, Image Processing, Security and Computer Vision, pp. 608-618, Springer, Cham.

[7] Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing.

[8] Pramerdorfer, C., & Kampel, M. (2016). Facial expression recognition using convolutional neural networks: state of the art. arXiv preprint arXiv:1612.02903.

[9] Facial Recognition Technology: Evolution and Applications "https://www.aiiottalk.com/facial-recognition-technology-evolution-and-applications/"

[10] C. Munteanu and A. Rosa, ‘‘Gray-scale image enhancement as an automatic process driven by evolution,’’ IEEE Trans. Syst., Man Cybern., B, Cybern., vol. 34, no. 2, pp. 1292–1298, Apr. 2004.

[11] L. Liu, P. Fieguth, Y. Guo, X. Wang, and M. Pietikäinen, ‘‘Local binary features for texture classification: Taxonomy and experimental study,’’ Pattern Recognit., vol. 62, pp. 135–160, Feb. 2017

[12] U. R. Acharya, Y. Hagiwara, J. E. W. Koh, J. H. Tan, S. V. Bhandary, A. K. Rao, and U. Raghavendra, ‘‘Automated screening tool for dry and wet age-related macular

degeneration (ARMD) using pyramid of histogram of oriented gradients (PHOG) and nonlinear features,'' J. Comput. Sci., vol. 20, pp. 41–51, May 2017.

[13] C. Turan and K.-M. Lam, ''Histogram-based local descriptors for facial expression recognition (FER): A comprehensive study,'' J. Vis. Commun. Image Represent., vol. 55, pp. 331–341, Aug. 2018.

[14] Facial Emotion Recognition: A Brief Review, Illiana Azizan, K. Fatimah, Universiti Putra Malaysia, International Conference on Sustainable Engineering, Technology and Management (ICSETM -2018), Dec. 20, 2018, Negeri Sembilan, Malaysia.

[15] Ekman, P., W. Friesen, Measuring Facial Movement, Environmental Psychology and Noverbal Behavior 1(1), Fall, 1976.

[16] Recognizing Action Units for Facial Expression Analysis, Ying-li Tian, Member, IEEE, Takeo Kanade, Fellow, IEEE, and Jeffrey F. Cohn, Member, IEEE; IEEE Trans Pattern Anal Mach Intell. 2001 Feb; 23(2): 97–115.

[17] Examples of Actions Units (AU) from Facial Action Coding System "https://link.springer.com/article/10.3758/s13423-017-1338-0/figures/1"

[18] BIOLOGICALLY VS. LOGIC INSPIRED ENCODING OF FACIAL ACTIONS AND EMOTIONS IN VIDEO, M.F. Valstar and M. Pantic,

[19] S. Deshmukh, M. Patwardhan, and A. Mahajan, "Survey on Real-Time Facial Expression Recognition Techniques," IET Biom., pp. 1-9, 2015.

[20] Viola, P., & Jones, M. (2001, December). Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001 (Vol. 1, pp. I-I). Ieee.

[21] Soo, S. (2014). Object detection using Haar-cascade Classifier. Institute of Computer Science, University of Tartu, 2(3), 1-12.

[22] Yang, Z., Li, M., & Ai, H. (2006, August). An experimental study on automatic face gender classification. In 18th International Conference on Pattern Recognition (ICPR'06) (Vol. 3, pp. 1099-1102). IEEE.

[23] Abuzneid, M. A., & Mahmood, A. (2018). Enhanced human face recognition using LBPH descriptor, multi-KNN, and back-propagation neural network. IEEE access, 6, 20641-20651.

[24] Mohammadi, M., Al-Azab, F., Raahemi, B., Richards, G., Jaworska, N., Smith, D., ... & Knott, V. (2015). Data mining EEG signals in depression for their diagnostic value. BMC medical informatics and decision making, 15(1), 1-14.

[25] Sharma, A. K., Kumar, U., Gupta, S. K., Sharma, U., & LakshmiAgrwal, S. (2018). A survey on feature extraction technique for facial expression recognition system. In 2018 4th International Conference on Computing Communication and Automation (ICCCA), pp. 1-6, IEEE.

[26] J. J. Pao, "Emotion Detection through Facial Feature Recognition," p. 6, 2018.

[27] Gao, Y., & Leung, M. K. (2002). Face recognition using line edge map. IEEE transactions on pattern analysis and machine intelligence, 24(6), 764-779.

[28] Ahmadian, A., & Mostafa, A. (2003, September). An efficient texture classification algorithm using Gabor wavelet. In Proceedings of the 25th annual international conference of the IEEE engineering in medicine and biology society (IEEE Cat. No. 03CH37439) (Vol. 1, pp. 930-933). IEEE.

[29] Kumar, Y., & Sharma, S. (2017). A systematic survey of facial expression recognition techniques. In 2017 international conference on computing methodologies and communication (ICCMC), pp. 1074-1079, IEEE

[30] Harshitha, S., Sangeetha, N., Shirly, A. P., & Abraham, C. D. (2019). Human facial expression recognition using deep learning technique. In 2019 2nd International Conference on Signal Processing and Communication (ICSPC), pp. 339-342, IEEE.

[31] Wu, T., Fu, S., & Yang, G. (2012). Survey of the facial expression recognition research. In International Conference on Brain Inspired Cognitive Systems, pp. 392-402, Springer, Berlin, Heidelberg.

[32] About Support Vector Machine Algorithm and its types details: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

[33] Hernández-Orallo, J. (2017). Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. Artificial Intelligence Review, 48(3), 397-447.

[34] Montalbo, F. J. P., & Alon, A. S. (2021). Empirical Analysis of a Fine-Tuned Deep Convolutional Model in Classifying and Detecting Malaria Parasites from Blood Smears. KSII Transactions on Internet and Information Systems (TIIS), 15(1), 147-165..

[35] Understanding the receptive field of deep convolutional networks: https://theaisummer.com/receptive-field/

[36] S. Linnainmaa, “Taylor expansion of the accumulated rounding error,” BIT Numer. Math., vol. 16, no. 2, pp. 146–160, 1976.

[37] The Most Intuitive and Easiest Guide for Convolutional Neural Network " https://towardsdatascience.com/the-most-intuitive-and-easiest-guide-for-convolutional-neural-network-3607be47480"

[38] Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. sensors, 18(2), 401.

[39] Huang, Y., Chen, F., Lv, S., & Wang, X. (2019). Facial expression recognition: A survey. Symmetry, 11(10), 1189.

[40] Papageorgiou, C. P., Oren, M., & Poggio, T. (1998). A general framework for object detection. In Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271), pp. 555- 562, IEEE

[41] Viola, P., & Jones, M. J. (2004). Robust real-time face detection. International journal of computer vision, 57(2), pp. 137-154.

[42] Weilong Chen, MengJooEr, Shiqian Wu (2006), "Illumination Compensation and Normalization forRobust Face Recognition Using Discrete Cosine Transform in Logarithm Domain", IEEE transactions on systems, man and cybernetics— part b: cybernetics, Vol. 36(2) pp.458-466.

[43] Owusu, E., Zhan, Y., & Mao, Q. R. (2014). A neural-AdaBoost based facial expression recognition system. Expert Systems with Applications, 41(7), pp. 3383-3390.

[44] Biswas, S., & Sil, J. (2015). An efficient expression recognition method using contourlet transform. In Proceedings of the 2nd International Conference on Perception and Machine Intelligence, pp. 167-174.

[45] Ji, Y., & Idrissi, K. (2012). Automatic facial expression recognition based on spatiotemporal descriptors. Pattern Recognition Letters, 33(10), pp. 1373-1380

[46] Zhang, L., Tjondronegoro, D., & Chandran, V. (2014). Random Gabor based templates for facial expression recognition in images with facial occlusion. Neurocomputing, 145, pp. 451- 464

[47] Happy, S. L., & Routray, A. (2014). Automatic facial expression recognition using features of salient facial patches. IEEE transactions on Affective Computing, 6(1), pp. 1-12.

[48] Happy, S. L., George, A., & Routray, A. (2012). A real time facial expression classification system using local binary patterns. In 2012 4th International conference on intelligent human computer interaction (IHCI) (pp. 1-5). IEEE.

[49] Ghimire, D., Jeong, S., Lee, J., & Park, S. H. (2017). Facial expression recognition based on local region specific features and support vector machines. Multimedia Tools and Applications, 76(6), pp. 7803-7821.

[50] Ghimire, D., & Lee, J. (2013). Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. Sensors, 13(6), pp. 7714- 7734

[51] Bhadu, A., Kumar, V., Shekhawat, H. S., & Tokas, R. (1956). An improved method of feature extraction technique for facial expression recognition using Adaboost neural network. International Journal of Electronics and Computer Science Engineering (IJECSE) Volume, 1, pp. 1112-1118.

[52] Zhao, G., Huang, X., Taini, M., Li, S. Z., & PietikäInen, M. (2011). Facial expression recognition from near-infrared videos. Image and Vision Computing, 29(9), pp. 607-619.

[53] Shen, P., Wang, S., & Liu, Z. (2013). Facial expression recognition from infrared thermal videos. In Intelligent Autonomous Systems 12, pp. 323-333, Springer, Berlin, Heidelberg.

[54] Szwoch, M., & Pieniążek, P. (2015). Facial emotion recognition using depth data. In 2015 8th International Conference on Human System Interaction (HSI), pp. 271-277, IEEE.

[55] Wei, W., Jia, Q., & Chen, G. (2016). Real-time facial expression recognition for affective computing based on Kinect. In 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA), pp. 161-165, IEEE.

[56] Sohail, A. S. M., & Bhattacharya, P. (2007). Classification of facial expressions using k-nearest neighbor classifier. In International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications, pp. 555-566, Springer, Berlin

[57] Wang, X. H., Liu, A., & Zhang, S. Q. (2015). New facial expression recognition based on FSVM and KNN. Optik, 126(21), pp. 3132-3134.

[58] Valstar, M., Patras, I., & Pantic, M. (2004). Facial action unit recognition using temporal templates. In RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759), pp. 253-258, IEEE.

[59] Chen, L., Zhou, C., & Shen, L. (2012). Facial expression recognition based on SVM in Elearning. Ieri Procedia, 2, pp. 781-787.

[60] Michel, P., & El Kaliouby, R. (2003). Real time facial expression recognition in video using support vector machines. In Proceedings of the 5th international conference on Multimodal interfaces, pp. 258-264.

[61] Tsai, H. H., & Chang, Y. C. (2018). Facial expression recognition using a combination of multiple facial features and support vector machine. Soft Computing, 22(13), pp. 4389-4405

[62] Hsieh, C. C., Hsih, M. H., Jiang, M. K., Cheng, Y. M., & Liang, E. H. (2016). Effective semantic features for facial expressions recognition using SVM. Multimedia Tools and Applications, 75(11), pp. 6663-6682.

[63] Mahersia, H., & Hamrouni, K. (2015). Using multiple steerable filters and Bayesian regularization for facial expression recognition. Engineering Applications of Artificial Intelligence, 38, pp. 190-202.

[64] Mahersia, H., & Hamrouni, K. (2015). Using multiple steerable filters and Bayesian regularization for facial expression recognition. Engineering Applications of Artificial Intelligence, 38, pp. 190-202.

[65] Neggaz, N., Besnassi, M., & Benyettou, A. (2010). Application of improved AAM and probabilistic neural network to facial expression recognition. Journal of Applied Sciences(Faisalabad), 10(15), pp. 1572-1579

[66] Lopes, A. T., de Aguiar, E., De Souza, A. F., & Oliveira-Santos, T. (2017). Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recognition, 61, pp. 610-628.

[67] Agrawal, A., & Mittal, N. (2020). Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy. The Visual Computer, 36(2), pp. 405-412.

[68] Jain, D. K., Shamsolmoali, P., & Sehdev, P. (2019). Extended deep neural network for facial emotion recognition. Pattern Recognition Letters, 120, pp. 69-74.

[69] Liu, P., Han, S., Meng, Z., & Tong, Y. (2014). Facial expression recognition via a boosted deep belief network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1805-1812.

[70] Burkert, P., Trier, F., Afzal, M. Z., Dengel, A., & Liwicki, M. (2015). Dexpression: Deep convolutional neural network for expression recognition. arXiv preprint arXiv:1509.05371.

[71] Khaireddin, Y., & Chen, Z. (2021). Facial Emotion Recognition: State of the Art Performance on FER2013. arXiv preprint arXiv:2105.03588.

[72] Xu, Q., Wang, C., & Hou, Y. (2021, January). Attention Mechanism and Feature Correction Fusion Model for Facial Expression Recognition. In 2021 6th International Conference on Inventive Computation Technologies (ICICT) (pp. 786-793). IEEE.

[73] Liu, J., Wang, H., & Feng, Y. (2021). An End-to-End Deep Model With Discriminative Facial Features for Facial Expression Recognition. IEEE Access, 9, 12158-12166.

[74] Li, S., Li, W., Wen, S., Shi, K., Yang, Y., Zhou, P., & Huang, T. (2021). Auto-FERNet: A Facial Expression Recognition Network with Architecture Search. IEEE Transactions on Network Science and Engineering.

[75] Mohan, K., Seal, A., Krejcar, O., & Yazidi, A. (2020). Facial expression recognition using local gravitational force descriptor-based deep convolution neural networks. IEEE Transactions on Instrumentation and Measurement, 70, 1-12.

[76] Pham, L., Vu, T. H., & Tran, T. A. (2021, January). Facial Expression Recognition Using Residual Masking Network. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 4513-4519). IEEE.

[77] Y. Tang, “Deep learning using linear support vector machines,” arXiv preprint arXiv:1306.0239, 2013

[78] Khalid, M., Baber, J., Kasi, M. K., Bakhtyar, M., Devi, V., & Sheikh, N. (2020, July). Empirical Evaluation of Activation Functions in Deep Convolution Neural Network for Facial Expression Recognition. In 2020 43rd International Conference on Telecommunications and Signal Processing (TSP) (pp. 204-207). IEEE.

[79] Hasani, B., Negi, P. S., & Mahoor, M. (2020). BReG-NeXt: Facial affect computing using adaptive residual networks with bounded gradient. IEEE Transactions on Affective Computing.

[80] Chhikara, P., Singh, P., Tekchandani, R., Kumar, N., & Guizani, M. (2020). Federated learning meets human emotions: A decentralized framework for human–computer interaction for iot applications. IEEE Internet of Things Journal, 8(8), 6949-6962.

[81] Hua, W., Dai, F., Huang, L., Xiong, J., & Gui, G. (2019). HERO: Human emotions recognition for realizing intelligent Internet of Things. IEEE Access, 7, 24321-24332.

[82] Pham, T. T. D., Kim, S., Lu, Y., Jung, S. W., & Won, C. S. (2019). Facial action units-based image retrieval for facial expression recognition. IEEE Access, 7, 5200-5207.

[83] Nguyen, D. H., Kim, S., Lee, G. S., Yang, H. J., Na, I. S., & Kim, S. H. (2019). Facial expression recognition using a temporal ensemble of multi-level convolutional neural networks. IEEE Transactions on Affective Computing.

[84] Li, M., Xu, H., Huang, X., Song, Z., Liu, X., & Li, X. (2018). Facial expression recognition with identity and emotion joint learning. IEEE Transactions on Affective Computing.

[85] Zeng, G., Zhou, J., Jia, X., Xie, W., & Shen, L. (2018, May). Hand-crafted feature guided deep learning for facial expression recognition. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) (pp. 423-430). IEEE.

[86] Ramalingam, S., & Garzia, F. (2018, October). Facial expression recognition using transfer learning. In 2018 International Carnahan Conference on Security Technology (ICCST) (pp. 1-5). IEEE.

[87] Xiang, J., & Zhu, G. (2017, July). Joint face detection and facial expression recognition with MTCNN. In 2017 4th international conference on information science and control engineering (ICISCE) (pp. 424-427). IEEE.

[88] Kankanamge, S., Fookes, C., & Sridharan, S. (2017, September). Facial analysis in the wild with LSTM networks. In 2017 IEEE International Conference on Image Processing (ICIP) (pp. 1052-1056). IEEE.

[89] Li, Z. (2017, December). A discriminative learning convolutional neural network for facial expression recognition. In 2017 3rd IEEE international conference on computer and communications (ICCC) (pp. 1641-1646). IEEE.

[90] Lu, P., Li, B., Shama, S., King, I., & Chan, J. H. (2017, November). Regularizing the loss layer of CNNs for facial expression recognition using crowdsourced labels. In 2017 21st Asia Pacific Symposium on Intelligent and Evolutionary Systems (IES) (pp. 31-36). IEEE.

[91] Mollahosseini, A., Chan, D., & Mahoor, M. H. (2016, March). Going deeper in facial expression recognition using deep neural networks. In 2016 IEEE Winter conference on applications of computer vision (WACV) (pp. 1-10). IEEE.

[92] Huang, Y., & Lu, H. (2016, December). Hybrid hypergraph construction for facial expression recognition. In 2016 23rd International Conference on Pattern Recognition (ICPR) (pp. 4142-4147). IEEE.

[93] Ng, H. W., Nguyen, V. D., Vonikakis, V., & Winkler, S. (2015, November). Deep learning for emotion recognition on small datasets using transfer learning. In Proceedings of the 2015 ACM on international conference on multimodal interaction (pp. 443-449).

[94] FER2013 dataset "https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data"

[95] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete expression dataset for action unit and emotion-specified expression. Proceedings of the Third International Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB 2010), San Francisco, USA, 94-101.

[96] Michael J. Lyons, Shigeru Akamatsu, Miyuki Kamachi, Jiro Gyoba.
Coding Facial Expressions with Gabor Wavelets, 3rd IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200-205 (1998).
http://doi.org/10.1109/AFGR.1998.670949

[97] Hawkins, D. M. (2004). The problem of overfitting. Journal of chemical information and computer sciences, 44(1), 1-12.

[98] Faridul, H. S., Pouli, T., Chamaret, C., Stauder, J., Trémeau, A., & Reinhard, E. (2014). A Survey of Color Mapping and its Applications. Eurographics (State of the Art Reports), 3(2), 1.

[99] Color Mapping in OpenCV " https://medium.com/@pragyatomar1611/color-mapping-in-opencv-637cdd50c603"

[100] Qassim, H., Verma, A., & Feinzimer, D. (2018, January). Compressed residual-VGG16 CNN model for big data places image recognition. In 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 169-175). IEEE.

[101] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, “Densely connected convolutional networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, vol. 1, no. 2,p. 3, 2017

[102] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, ‘‘AID: A benchmark data set for performance evaluation of aerial scene classification,’’ IEEE Trans. Geosci. Remote Sens., vol. 55, no. 7, pp. 3965–3981, Jul. 2017, doi: 10.1109/TGRS.2017.2685945.

[103] Alhichri, H., Alswayed, A. S., Bazi, Y., Ammour, N., & Alajlan, N. A. (2021). Classification of remote sensing images using EfficientNet-B3 CNN model with attention. IEEE Access, 9, 14078-14094.

[104 Tan, M., & Le, Q. V. (2021). Efficientnetv2: Smaller models and faster training. arXiv preprint arXiv:2104.00298.

[105] Aouayeb, M., Hamidouche, W., Soladie, C., Kpalma, K., & Seguier, R. (2021). Learning Vision Transformer with Squeeze and Excitation for Facial Expression Recognition. arXiv preprint arXiv:2107.03107.

[106] Meng, D., Peng, X., Wang, K., & Qiao, Y. (2019, September). Frame attention networks for facial expression recognition in videos. In 2019 IEEE International Conference on Image Processing (ICIP) (pp. 3866-3870). IEEE.

[107] Ding, H., Zhou, S. K., & Chellappa, R. (2017, May). Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017) (pp. 118-126). IEEE.

[108] Chen, Y., Wang, J., Chen, S., Shi, Z., & Cai, J. (2019, December). Facial motion prior networks for facial expression recognition. In 2019 IEEE Visual Communications and Image Processing (VCIP) (pp. 1-4). IEEE.

[109] Minaee, S., Minaei, M., & Abdolrashidi, A. (2021). Deep-emotion: Facial expression recognition using attentional convolutional network. Sensors, 21(9), 3046.

[110] Happy, S. L., & Routray, A. (2014). Automatic facial expression recognition using features of salient facial patches. IEEE transactions on Affective Computing, 6(1), 1-12.

[111] Minaee, S., Minaei, M., & Abdolrashidi, A. (2021). Deep-emotion: Facial expression recognition using attentional convolutional network. Sensors, 21(9), 3046.

[112] Shima, Y., & Omori, Y. (2018, August). Image augmentation for classifying facial expression images by using deep neural network pre-trained with object image database. In Proceedings of the 3rd International Conference on Robotics, Control and Automation (pp. 140-146).

[113] Liu, C., Hirota, K., Ma, J., Jia, Z., & Dai, Y. (2021). Facial Expression Recognition Using Hybrid Features of Pixel and Geometry. IEEE Access, 9, 18876-18889.

[114] Zheng, W., Zhou, X., Zou, C., & Zhao, L. (2006). Facial expression recognition using kernel canonical correlation analysis (KCCA). IEEE transactions on neural networks, 17(1), 233-238.

[115] Kommineni, J., Mandala, S., Sunar, M. S., & Chakravarthy, P. M. (2021). Accurate computing of facial expression recognition using a hybrid feature extraction technique. The Journal of Supercomputing, 77(5), 5019-5044.