Addressing Class Imbalance in Facial Emotion Recognition

by

Sarvenaz Ghafourian Bolori Mashhad
B.Sc., University of Shahid Rajaee, 2019

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Engineering

in the Department of Electrical and Computer Engineering

Addressing Class Imbalance in Facial Emotion Recognition

by

Sarvenaz Ghafourian Bolori Mashhad
B.Sc., University of Shahid Rajaee, 2019

Supervisory Committee

Dr. Amirali Baniasadi , Supervisor
(Department of Electrical and Computer Engineering)

Dr. T. Aaron Gulliver , Committee Member
(Department of Electrical and Computer Engineering)

ABSTRACT

The wide usage of computer vision in many perspectives has been attracted in the recent years. One of the areas of computer vision that has been studied is facial emotion recognition, which plays a crucial role in the interpersonal communication. This work demonstrates the advances could be made in this field. This work tackles the problem of intraclass variances in the face images of emotion recognition datasets. We test the system on an augmented dataset including CK+, EMOTIC, and KDEF dataset samples, which increase the intraclass variances in the face images of our dataset. The proposed method is based on SMOTETomek.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I am indebt to my supervisor, Dr. Amirali Baniasadi, for his patience, guidance, and support. I have benefited greatly from his wealth of knowledge. I am extremely grateful that he took me on as a student and continued to have faith in me over the years.

I would also like to thank my parents, for their unconditional, unequivocal, and endless support. They have always stood behind me, and this was no exception.I am grateful for my parents whose constant love and support keep me motivated and confident. My accomplishments and success are because they believed in me.

DEDICATION

I dedicate my dissertation work to my family. A special feeling of gratitude to my loving parents, whose words of encouragement and push for tenacity ring in my ears. My sister who has never left my side and is very special.

Last but not least, I want to dedicate this thesis to my maternal grandmother who has meant and continue to mean so much to me. Although she is no longer of this world, her memories continue to regulate my life.

Chapter 1

Introduction

1.1 Problem Statement

Since the human face plays an integral part in expressing a person's mental state, facial expression analysis is a significant research focus with numerous potential uses. Scientists from many areas like psychology, finance, marketing, and engineering have been greatly interested in this subject due to the practical benefits. Artificial intelligence is becoming more prevalent in many aspects of human life. The technologies are adapted to the needs of human beings, and artificial intelligence is what makes this adaptation between technology and humans possible. While it may come easy for most humans to process emotions without any extra effort, computers have struggled with the idea of recognizing them automatically for decades. This challenge is due to face appearance changes caused by pose variations, illumination variations, and camera quality and angle changes.

Research from different disciplines such as computer vision and machine learning focus on utilizing computers to categorize emotions exhibited by humans properly. Analyzing human facial expressions is part of this work. Specifically, we study the task of facial emotion recognition based on two deep learning models using our suggested dataset. We use various face images for seven emotions and propose how to improve the efficiency of emotion detection.

1.2 Emotions and Emotion Recognition

Emotions are related to mental states, thoughts, feelings, behavioral responses, and levels of happiness or dissatisfaction caused by neurophysiological changes. Emotions are often intertwined with mood, temperament, personality, creativity, and motivation. From a scientific viewpoint, emotions may be defined as "a pleasant or unpleasant experience associated with a certain pattern of physiological activity". Emotions have an impact on physiological, behavioral, and cognitive changes. Emotions' initial purpose was to promote adaptive behaviour, which may have aided in the transfer of genes through survival, reproduction, and kinship selection in the past.

Emotion research has grown over the past two decades, with contributions in many fields, including psychology, medicine, history, the sociology of emotion, and computer science. In addition, several ideas that have attempted to explain the origin, function, and other aspects of emotion have encouraged more in-depth research on the topic. The technique of recognizing human emotion is known as emotion recognition. People's ability to recognize other people's emotions varies greatly. The use of technology to assist humans in recognizing emotions is a relatively new study topic [13]. The study of biological features, which clearly indicate a person's emotional state, is the foundation of recognition.

Automatic emotion detection systems are now widely utilized in various areas, including health, the entertainment industry, smart home management systems, cybersecurity, and distant learning systems. However, according to an examination of the market for current means of recognizing an emotional state, the means that are based on the analysis of the parameters characterizing the geometry of the person's face have the most extensive distribution in general-purpose information systems.

The face is defined as the front portion of the human head, from above by the scalp's border, below by the corners and bottom edge of the lower jaw, and on the sides by the margins of the lower jaw's branches and the base of the auricles [14].

Because facial expressions connect to emotions, they are essential identifiers for human sentiments. Facial expression is a nonverbal means of expressing emotion in the majority of situations (about 55 percent of the time), and it may be used as a tangible evidence to determine whether someone is telling the truth or not. According to studies, studying facial expressions can influence how we perceive what is said and how a conversation flows. For example, 93 percent of communication can be attributed to an entity's emotion in a typical conversation. Thus, humans can

perceive emotions that are critical for effective communication.

Facial recognition is one of the factors involved in emotion recognition. Emotions, as previously said, are inherent characteristics of people that play a significant part in social communication [15, 16]. Humans show emotion in a variety of ways, including facial expressions [17], gestures, vocalizations, body language [18], and speech [19]. The most popular and well-researched field is expression analysis. The definitions of six basic emotions described by Eckman are:

- **Happiness:** Of all the various emotions, people aspire to happiness out of all kinds of emotions. Contentment, joy, fulfillment, satisfaction, and well-being are all familiar sensations associated with happiness. In terms of facial expression, this sort of feeling is sometimes represented by smiling [20].

- **Sadness:** An accompanying state of sadness is marked by disappointment, grief, hopelessness, indifference, and a depressed mood. Sadness can be expressed in a number of ways including crying, dampened mood, lethargy, and withdrawal from others.

- **Fear:** Fear is a strong emotion that helps people survive in some situations, and it usually trigger the fight response. Fear expressions in the face include widening the eyes and pushing the chin back. In addition, your muscles tense up, your heart rate and respiration speed up, and your mind becomes more alert, preparing your body to combat the threat.

- **Disgust:** Many factors can trigger feelings of disgust, including an unpleasant taste, sight, or smell. The emotion is believed to be a reaction to potentially dangerous or spoiled foods. Face expressions such as wrinkling the nose and curling the upper lip are examples of disgust.

- **Anger:** Anger consists of feelings of agitation, frustration, and animosity toward others. Anger, like fear, may trigger your body's reaction. Fretting or staring in the face is a common expression of anger.

- **Surprise:** Eckman also labels this emotion as one of the six fundamental types of human emotions. Surprise is the response of the body to something unexpected. It typically lasts only a few seconds. This feeling can be either positive, negative, or neutral. Expanding the eyes, raising the brows, and opening the mouth are all signs of surprise.

In computer vision and robotics, automated face expression recognition is a task. This is a new area of study, particularly in social signal processing and emotional computing fields. Automated facial expression recognition aims to detect and distinct facial expressions into corresponding emotion groups [21]. This topic has a broad range of applications, including entertainment, education, ecommerce, health, and security [22, 23]. Humans are geared towards understanding their counterparts' emotions and being intuitive about them in day-to-day social interaction. As for computers, however, determining the emotion is more challenging [24].

Our goal is to identify an individual's emotion from observing their facial expressions. First, cropped headshots are extracted using the FaceNet architecture. Second, the extracted face images from three different datasets are used as a single dataset for the transfer-learning task on VGG-16 and ResNet- 50. The model would output a probability distribution over the emotions (angry, disgust, neutral, sad, surprise, fear, happy) of the pictured individual. Thus, our work focuses on the following contribution:

1. A comprehensive analysis of popular emotion recognition datasets, such as CK+, EMOTIC, and KDEF. We describe how images are categorized in each dataset.

2. Creating a custom dataset consisting of the three above-mentioned datasets to cover a wide range of variations in face images. We explain how different images of our dataset are cropped to fit our criteria.

3. Improving class imbalance problem in the custom emotion recognition dataset over VGG-16 and ResNet-50. We show how SMOTETomek technique improves the distinction accuracy over VGG-16 and ResNet-50 models.

1.3 Motivation

In present-day society, computers play an increasingly important role in a variety of aspects of our lives. Traditional computer interfaces, such as keyboard and mouse, are no longer enough to meet human needs. As a result, establishing a natural connection between machines and humans is essential. In contrast to human-to-human contact, human-computer interaction has thus far been much less natural. Although, human-computer interaction cannot wholly replace face-to-face communication.

Emotional recognition and expression are the most critical prerequisites for computers. Similarly, they need to have the ability to perceive and comprehend the emotions of their human counterparts in order to develop a more genuine relationship [24]. Thanks to the rapid expansion of computers and the Internet, understanding people's emotions and reacting appropriately is becoming increasingly important to video games, human-computer interactions, and cognitive and affective computing applications. As a result, emotion detection and its applications are becoming increasingly popular in both scientific study and industry. On the other hand, Computers find it challenging to monitor or understand human emotion since emotion is complex in and of itself [24]. Furthermore, it is demonstrated that one of the most important disadvantages of contemporary neural network emotion detection methods employed in general-purpose information systems is the lack of recognition accuracy when distinctive interference is present.

Overfitting occurs when the Facial Emotion Recognition (FER) model is trained on imbalanced datasets [25], making the model less capable of performing FER tasks in real-world scenarios. As a result, overfitting due to a lack of data remains a problem for most FER systems. Thus, we are motivated to create an augmented dataset in this work to mitigate the overfitting problem and improving generalization.

Moreover, the intrinsic imbalances in the distribution of facial expression samples in the dataset, which is due to the nature of emotions, might degrade the system's overall performance [26]. Therefore, we are motivated to handle the fact that the number of collected facial images for the primary classes, such as happiness, is much larger than for the minor classes, such as fear.

1.4 Related Work

The traditional approach to emotion detection consists of a two-stage machine learning process, in which the

rst phase involves collecting characteristics from the pictures, and the second phase involves using a classi

er (such as an SVM, neural network, or random forest) to determine the emotions.

The histogram of oriented gradients (HOG) [27], local binary patterns (LBP) [28], Gabor wavelets [29], and Haar features [30] are some of the prominent hand-crafted features utilized for face emotion identification. The appropriate emotion is then assigned to the image using a classifier.

While these methods work for small datasets, they start showing their limits when applied to more complex datasets (with higher intraclass variances). In other words, there are some issues with face images when the face is partially visible [31].

The majority of contemporary computer vision research into recognizing people's emotional states is based on facial expression analysis [32]. Psychologists, Ekman and Friesen, identified six fundamental emotions and multiple methods for recognizing them. The Facial Action Coding System is used in several of these approaches. Action Units (AU) are a collection of unique localized movements of the face that encode facial emotion. This approach uses a set of distinct localized facial movements known as Action Units to represent facial emotion.

CNNs have been used in recent studies for emotion detection based on facial expression to recognize emotions and Action Units [32]. In response to the great success of deep learning and, in particular, CNNs (convolutional neural networks) for image classification and other vision challenges, a number of organisations have built deep learning-based facial expression recognition (FER) models [33]. Mollahosseini et al. showed that CNNs could recognize emotions accurately and achieve state-of-the-art results. The results are based on a zerobiased CNN on the expanded Cohn-Kanade dataset (CK+) and the Toronto Face Dataset (TFD) [34]. Mollahosseini also, suggested an FER neural network with two convolution layers, one max-pooling layer, and four inception layers, in each layer.

Aneja et al. [35] created a model of facial expressions for stylized animated characters using deep learning. Their training included a network that represented human expressions, and a network that represented animated faces. The loopy network was first proposed by Liu in [36], noting the importance of feedback of the weak classifiers. Instead of using a strong classifier, a loop of weaker classifiers are used for emotion detection. This network is called a loopy network. They used their Boosted Deep Belief Network (BDBN) over CK+ and JAFFE datasets to achieve a higher accuracy.

In addition to determining the face characteristics, some studies [37] detect fundamental emotions using the position of shoulders. Schindler et al. [38] used a limited dataset of non-spontaneous postures obtained under controlled conditions to detect the six primary emotions. Rather than identifying emotion categories, some more recent research on facial expression [39] employs the VAD (Valence, Arousal, Dominance) Emotional State Model's continuous dimensions to describe emotions [40].

It should be noted that the majority of the past research is based on widely used facial expression recognition datasets, such as FER2013, the extended Cohn-Kanade

(CK+), and the Japanese Female Facial Expression dataset (JAFFE). These datasets consist of frontal face images, and the photos lack any contextualized backgrounds and have fewer differences, such as spectacles or face masks. This makes the facial action units detection easier. However, we expect our model to perform on more challenging images as well. Images consisting of illumination, pose, occlusion, and low resolution are considered challenging images.

Despite the fact that all previous studies on emotion detection have improved considerably, achieving a higher accuracy rate in this field has always been a goal because it is used in susceptible sectors such as security. We attempt to solve this issue by offering a customized dataset for achieving greater accuracy on an augmented, more challenging face image dataset and model it on two deep learning algorithms.

1.5 Agenda

Chapter 1 contains a problem statement that this dissertation will prove, followed by introducing emotion and emotion recognition concepts, motivation, and related works.

Chapter 2 introduces the face recognition concept and continuously describes in detail different segments of face recognition. The rest of the chapter is an illustration of some main facial recognition system challenges.

Chapter 3 explains the research, the algorithms involved, and the new dataset processing.

Chapter 4 is where different methodologies for the problem of face recognition are fully described. The first part includes details of the Neural Network and its multiple layers. The chapter is followed by the convolution size calculation, which is helpful to determine the size of different layers in the models analyzed in this work. The rest of the chapter is about the full description of the Residual Network and VGG Network that our dataset is testified on.

Chapter 5 includes the evaluation of the data presented above and the solution for the dataset challenge to show how much more accurate the new approach is.

Chapter 6 contains a restatement of the problem and the future work of the dissertation.

Chapter 2

Face Recognition

The challenge of recognizing and validating persons in an image by their faces is known as face recognition. The task is not difficult for humans, even in bright or dim light or with their faces obscured by obstacles. However, it has sill remained a complex computer vision challenge. Face recognition is sometimes defined as a four-step process that begins with face detection, then moves on to face alignment, feature extraction, and ultimately face identification (Fig 2.1).

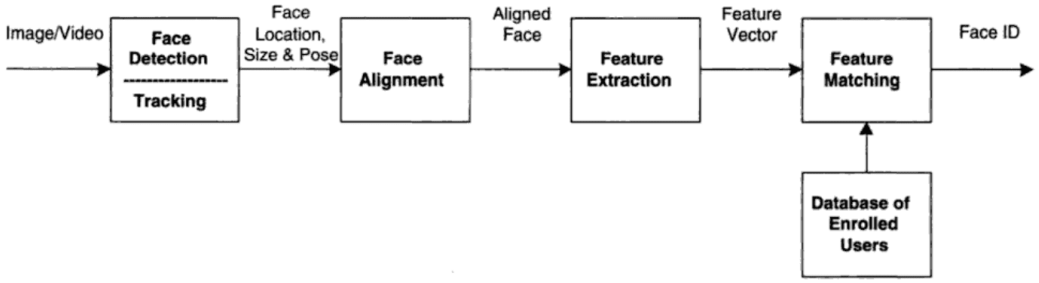

Figure 2.1: Face recognition processing flow [2]

1. **Face Detection:** A rounded box is drawn around one or more faces within the image.
2. **Face Alignment:** The face can be normalized to match the geometry and photometric of the database.
3. **Feature Extraction:** The job of facial recognition requires extracting facial characteristics.

4. **Face Recognition:** Perform matching of the face against one or more known faces in a prepared database.

2.1 Face Detection Task

Face detection is the challenge of identifying and localizing one or more faces in a picture in computer vision. Localization is the process of determining the extent of a person's face, usually by drawing a bounding box around it. Whereas, finding the coordinate of a person's face in a photograph is the process of determining the coordinate of a person's face in the image.

Face detection in photographs is simple for humans, but it has historically proven difficult for computers due to the dynamic aspect of faces. Faces, for example, must be recognized independently of their orientation or angle, light levels, clothes, accessories, hair color, facial hair, cosmetics, age, and so on.

2.2 Face Alignment Task

The effort of detecting the geometric structure of faces in digital pictures and attempting to achieve a canonical alignment of the face using translation, scale, and rotation is known as face alignment. This method is used in computer vision to identify the shape of a face in a digital picture. It automatically derives the form of facial components such as eyes and nose based on the location and size of a face. By iteratively modifying deformable models that remember previous face shape or appearance, face alignment applications can take low-level visual evidence into account and locate a face in a picture [41].

2.3 Feature Extraction Task

Feature extraction is used in machine learning, pattern recognition, and image processing to create derived values (features) that are meant to be valuable and non-redundant, easing future learning and generalization stages, and in certain circumstances, leading to improve human interpretations. Dimensionality reduction is linked to feature extraction.

An algorithm's input data can be reduced to a smaller collection of characteristics

(also referred to as a feature vector) if the data is too large to analyze and is suspected to be redundant (for example, the exact measurement in feet and meters or the same picture given as pixels). Selecting one or more subsets of the original features is the process of feature selection (Fig 2.2). A reduced representation of the input data should contain the essential information from the input data, enabling the intended job to be completed rather than using the complete data set.

Feature extraction is the process of minimizing the number of resources needed to explain a large amount of data. One of the most significant issues with doing sophisticated data analysis is the large number of variables involved. A high number of variables necessitates a lot of memory and processing resources. It can also lead a classification algorithm to overfit training data and fail to generalize to new samples. Feature extraction is a wide term that relates to methods for creating combinations of variables with enough precision [42].

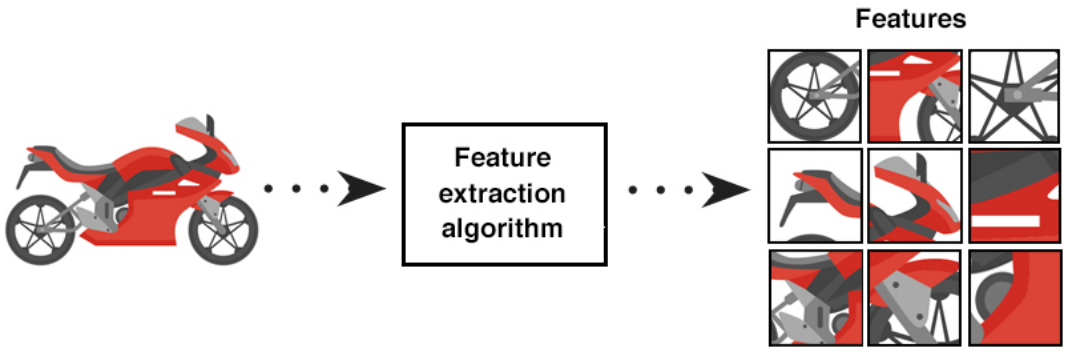

Figure 2.2: Feature extraction process [3]

2.4 Face Recognition Task

The job of identifying a face in a picture or video image against a pre-existing database of faces is known as facial recognition. It starts with detection, which involves identifying human faces from other things in the picture, then moving on to face recognition. Referring to the book, "Hand book of Face Recognition", face recognition consists of two primary modalities:

- **Face Verification:** An exact match between a face and an identified identity (e.g., Is this the right person?).

- **Face Identification:** Matching a given face with a database of known faces (i.e., who is this person?).

2.5 Facial Recognition System Challenges

2.5.1 Illumination

Light fluctuations are referred to as illumination. Automatic facial recognition is substantially compromised by slight variations in lighting and can significantly impact the results. If the light changes and the same person is photographed with the same sensor and in nearly the same facial expression and position, the results might be considerably different. The facial appearance is significantly altered by illumination. Researchers found that the difference between two identical faces photographed under various illuminations is more significant than between two distinct faces photographed under the same lighting [4].

Figure 2.3: An example of illumination [4]

2.5.2 Pose

Pose changes are pretty sensitive to facial recognition systems. When a person's head moves and their viewing angle changes, the posture of their face changes. As a result of head motions and differing camera angles, intraclass variances in face appearance result in considerably reduced automated face recognition rates. When the rotation angle is increased, identifying the natural face becomes more difficult [4].

Figure 2.4: An example of pose variation [4]

2.5.3 Occlusion

Occlusion refers to a blocking of one or more portions of the face, preventing the entire face from being used as an input picture. In a face recognition system, occlusion is one of the most challenging problems to solve. Real-world scenarios commonly cause this because of beards, mustaches, and using an accessories like sunglasses. This diversity makes computerized facial recognition a complex problem to crack [4].

Figure 2.5: An example of occlusion [4]

2.5.4 Low Resolution

Any typical picture should have a resolution of 16x16 pixels. Low-resolution images are those that have a resolution of less than 16*16 pixels. Small-scale standalone cameras, such as CCTV cameras in streets, ATM cameras, and supermarket security cameras, can be used to capture these low-resolution pictures [43]. However, these cameras can only capture a human face's central 16x16 face region since they do not come close enough to it. Such a low-quality image loses most of its pixels, so it does not provide much information. As a result, the process of identifying faces may be pretty tricky [4].

Chapter 3

Data Processing

3.1 FaceNet

Face recognition is suggested in this paper using a novel network because the pictures in our dataset mainly consist of face images from various angles, faces with occlusion, various backgrounds, and faces under variable light.

Using FaceNet [44], an embedded face recognition and clustering network is the recommended approach to detect faces. Face recognition has undergone considerable improvements in recent years [45, 46]. A new system called FaceNet learns a mapping from photographs of faces to a compact Euclidean space, in which distances directly correlate with measures of face similarity. After this space has been created, tasks like face recognition, verification, and clustering may be readily done using traditional approaches using FaceNet embeddings as feature vectors [44].

FaceNet is a Google-developed facial recognition system that obtained state-of-the-art results on various face identification benchmark datasets in 2015 [44]. Due to the availability of pre-trained models and numerous third-party open-source implementations of the FaceNet system, it may be utilized widely. FaceNet may be used to extract high-quality characteristics from faces, referred to as face embeddings, which can subsequently be used to train a face recognition system.

The model is a triplet loss function-trained deep convolutional neural network that promotes vectors for the same identity to grow more similar (lower distance). In contrast, vectors for different identities are intended to become less similar (larger distance).

The focus on training a model to produce embeddings directly rather than ex-

tracting them from an intermediary layer of a model was a significant breakthrough in FaceNet. According to what the publisher of this article has said: "Rather than using an intermediary bottleneck layer as in prior deep learning techniques, we employ a deep convolutional network trained to optimize the embedding itself directly" [44].

The researchers employ triplets of approximately aligned face patches produced by a new three-fold online mining technique to train. The advantage of this method is considerably more representational: "we obtain state-of-the-art facial recognition performance with only 128 bytes per face" [44].

In this work, the FaceNet architecture is applied on to the set of images with different backgrounds and challenges. As an example, there might be a image of a kid hiding behind a tree and the face is visible partially. The system is tasked to detect the face in this image by analysing each segment of the image and match them with face detection standards.

3.2 Dataset Preparation

Several research studies suggest that static human faces are best suited to test the ability to detect facial emotion [47]. Unsurprisingly, there are several verified picture databases showing face expressions accessible in the literature [47].

In this work, three different databases, each with various features, are augmented to develop a large dataset. The datasets used to create our specific dataset are:

- EMOTIC Dataset
- KDEF Dataset
- CK+ Dataset

3.2.1 EMOTIC Dataset

Emotion in context (EMOTIC) [1] is a dataset of images with people in real environments, annotated with their apparent emotions. Since the photos are collected from real-life settings, the images' variations are higher than other common datasets, such as CK+ and KDEF. This dataset includes facial occlusion (usually with a hand), partial faces, low-contrast images, and eyeglasses.

An extended list of 26 emotion categories is defined in this dataset to annotate the images, combined with three standard continuous dimensions: Valence, Arousal, and Dominance. Rather than recognizing emotion categories, several new studies on facial expression employ the aspects of the VAD Emotional State Model to depict emotions. The VAD model describes emotions using three numerical dimensions (Fig 3.1 and Table 3.1):

- **Valence (V):** A scale evaluates how pleasant or pleasant a feeling is, from negative to positive [1].
- **Arousal (A):** It is a scale that assesses a person's level of agitation, ranging from nonactive/calm to agitated/ready to act [1].
- **Dominance (D):** It is a scale that evaluates a person's amount of control over a situation, ranging from submissive/non-control to dominant/in-control [1].

The pictures in the EMOTIC dataset are mostly from well-known datasets such as MSCOCOC [48] and ADE20K [49]. Manually downloading photos of various themes, locations, events, and circumstances were also done using the Google search engine to gather the samples. The EMOTIC database consists of 18316 images with 23788 people annotated.

Moreover, this dataset uses two methods to represent emotions:

- **Discrete Categories:** A total of 26 discrete emotional categories are identified, covering a wide variety of emotional states.
- **Continuous Dimensions:** The VAD Emotional State Model is also used to express emotions. The continuous dimension annotations are on a scale of 1 to 10.

Combining categories and continuous dimensions during the training stage results in a more robust system for recognizing emotional states.

3.2.2 KDEF Dataset

The Karolinska Directed Emotional Faces (KDEF) [5] is one of the most widely used human facial expressions databases. The Karolinska Directed Emotional Faces (KDEF) is a collection of 4900 photographs depicting human face emotions. There are 70 people in the photo collection, each with a different emotional expression.

1. Peace: well being and relaxed; no worry; having positive thoughts or sensations; satisfied
2. Affection: fond feelings; love; tenderness
3. Esteem: feelings of favorable opinion or judgment; respect; admiration; gratefulness
4. Anticipation: state of looking forward; hoping on or getting prepared for possible future events
3 **5. Engagement:** paying attention to something; absorbed into something; curious; interested
6. Confidence: feeling of being certain; conviction that an outcome will be favorable; encouraged; proud
7. Happiness: feeling delighted; feeling enjoyment or amusement
8. Pleasure: feeling of delight in the senses
9. Excitement: feeling enthusiasm; stimulated; energetic
10. Surprise: sudden discovery or something unexpected
11. Sympathy: state of sharing others emotions, goal or troubles; supportive; compassionate
12. Doubt/Confusion: difficulty to understand or decide; thinking about different options
13. Disconnection: feeling not interested in the main event of the surrounding; indifferent; bored; distracted
14. Fatigue: weariness; tiredness; sleepy
15. Embarrassment: feeling ashamed or guilty
16. Yearning: strong desire to have something; jealous; envious; lust
17. Disapproval: feeling that something is wrong or reprehensible; contempt; hostile
18. Aversion: feeling disgust, dislike, repulsion; feeling hate
19. Annoyance: bothered by something or someone; irritated; impatient; frustrated
20. Anger: intense displeasure or rage; furious; resentful
21. Sensitivity: feeling of being physically or emotionally wounded; feeling delicate or vulnerable
22. Sadness: feeling unhappy, sorrow, disappointed, or discouraged
23. Disquietment: nervous; worried; upset; anxious; tense; pressured; alarmed
24. Fear: feeling suspicious or afraid of danger; threat, evil or pain; horror
25. Pain: physically suffering
26. Suffering: psychological or emotional pain; distressed; anguished

Table 3.1: Proposed emotion categories with definitions [1].

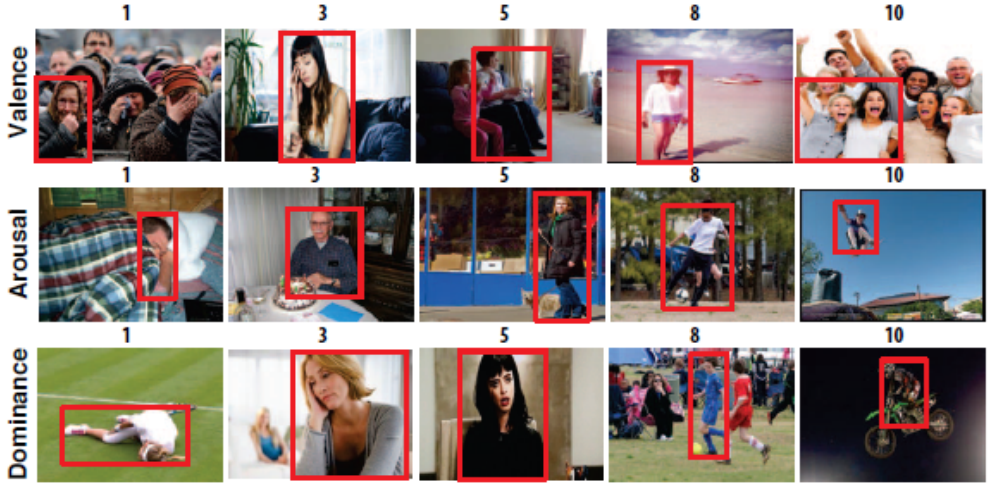

Figure 3.1: Examples of images from the EMOTIC dataset with different scores of Valence, Arousal, and Dominance [5].

Each emotion is examined from five distinct perspectives. All of the participants were given written instructions ahead of time [50]. These directions included a list of the seven various emotions they were supposed to make throughout the photo session. Before going to the photo session, the subject was instructed to practice the various expressions for an hour. It was stressed that the topic should strive to elicit the feeling being communicated, and - While keeping a natural method of expressing their feelings - aim to make the expression strong and obvious [5].

All of the participants wore unique grey T-shirts. They were around three meters away from the camera. Adjusting the camera position until the person's eyes and mouth were at particular, pre-defined vertical and horizontal places on the camera's grid screen, each subject's absolute distance was modified. The lights were adjusted to produce a gentle indirect light on both sides of the face, evenly dispersed. As part of the first series of photographs (series one), the individuals were photographed one at a time in different expressions. The second series of photos was taken (series two) in various poses and perspectives. Furthermore, an essential advantage of this simple rating method is that the instruction is straightforward. This may help the participants focus exclusively on their specific emotions for their intensity and arousal ratings.

3.2.3 CK+ Dataset

In 2000, the Cohn-Kanade (CK) database [51] was released for promoting research into automatically detecting individual facial expressions. Since then, the CK database has grown in popularity as a test-based for algorithm creation and assessment. Three constraints have become apparent at this time:

1) AU codes are adequately verified, but emotion labels are not because they correspond to what was requested rather than what was really done.

2) A lack of a single performance criterion against which new algorithms may be judged.

3) There are no standard protocols for shared databases.

As a result, the CK database has been utilized for both AU and emotion identification (despite the fact that the latter's labels have yet to be verified), there is no comparison with benchmark methods, and meta-analyses are problematic due to the usage of random subsets of the original database. We offer the Extended Cohn-Kanade (CK+) database to solve these and other problems. The number of sequences has grown by 22%, while the number of subjects has increased by 27%. The original CK dataset has been supplemented with 107 new sequences and 26 new individuals. Each sequence's target expression has been completely FACS coded, and emotion labels have been updated and confirmed. Each sequence's peak expression has been completely FACS coded, and emotion labels have been updated and verified using the FACS Investigators Guide [52], with a visual examination by emotion researchers confirming this.

The peak expression for each sequence is fully FACS coded and emotion labels have been revised and validated concerning the FACS Investigators Guide [52] confirmed by visual inspection by emotion researchers.

Chapter 4

Convolution Neural Network

4.1 Neural Network

As the name implies, a neural network is a collection of algorithms that tries to discover hidden connections in data by mimicking how the human brain works. In this respect, neural networks relate to systems of neurons, either biological or artificial. Furthermore, because neural networks can adapt to changing input, they can produce the best possible outcome without requiring the output criteria to be redesigned.

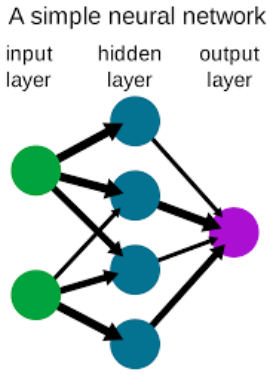

Figure 4.1: A simple neural network [6]

A neural network is similar to the neural network in the human brain. In a neural network, a "neuron" is a mathematical function that gathers and categorizes input using a specified design. The network closely resembles curve fitting and regression analysis, two statistical approaches. Layers of interconnected nodes make up a neural network. Each node is a perceptron, which works in the same way as multiple linear

regression. The perceptron converts the signal from a multiple linear regression into a nonlinear activation function.

The human brain is comprises neurons, which are nerve cells that transmit and interpret information from our senses. A network of nerves is formed when several of these nerve cells are grouped in our brain. Electrical impulses, or excitement, are sent via these nerves from one neuron to the next. Dendrites receive electrical impulses from resting terminal buttons or synapses of neighboring neurons. Dendrites transmit the impulse to the nerve cell's nucleus, also known as the soma. The electrical impulse is handled here before being sent to the axon. The axon is a longer dendritic branch that transports the impulse from the soma to the synapse. The impulse is subsequently sent to the dendrites of the second neuron through the synapse.

As a result, the human brain develops a complicated network of neurons. Machine learning methods employ the same notion of a network of neurons. The neurons are produced artificially on a computer in this example. An artificial neural network is created by connecting numerous of these artificial neurons. An artificial neuron works the same way as a neuron in our brain (Fig 4.2).

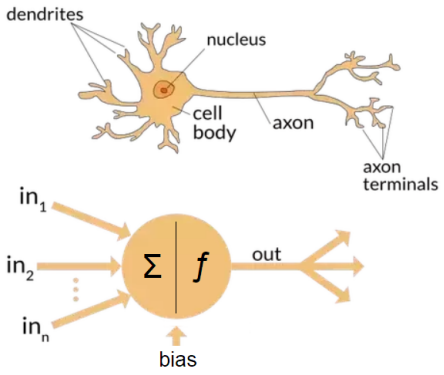

Figure 4.2: A network on neurons in the human body and a neural network in machine learning [7].

Each neuron in the network has a link that allows data to flow across it. The flow of data is controlled by a certain weight assigned to each link (Fig 4.3).

4.1.1 Convolutional Layer

An essential component of CNN architecture, the convolution layer extracts features using a combination of linear and nonlinear processes, such as convolution and activation functions [53].To put it another way, this layer's purpose is to extract valid

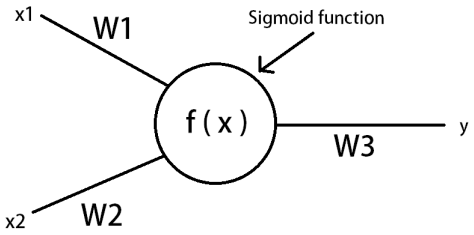

Figure 4.3: Flow of data calculation in a neural network

features from input images. This layer is applied to its input to generate the output which is called a feature map. The ultimate performance of the convolution layers is formed by combining all of these convolved maps.

4.1.2 Pooling Layer

A pooling layer performs a standard down-sampling operation on the feature maps, reducing their dimensionality to prevent the system runs out of memory. In other words, this layer's purpose is to isolate feature maps by summing or averaging their values across the convolved feature maps. The result is a single value for each square segment of the feature map [53]. This layer operates by sliding the pooling filter over the convolved feature with a larger convolved map than the pooling filter.

Max Pooling Layer

One of the most common types of pooling operation is the max-pooling layer, which takes square segments from input feature maps and pick the largest value in each segment. Usually, a max-pooling filter with a size 2×2 stride is used. As a result, the in-plane dimension of feature maps is reduced by a factor of two [53].

Average Pooling Layer

This layer is another common method of downsampling, while the depth is maintained [54]. It works by taking the average of all the components of each square segment in the feature map. The following are some of the benefits of using global average pooling:

- It lowers the number of parameters that may be learned.
- It allows the CNN to take inputs of varying sizes [53].

4.1.3 Fully Connected Layer

The goal of this layer is to convert a pooled feature map from a 2D structure to a 1D array of integers. This means that the final convolution or pooling layer's output feature maps are generally flattened. This layer inputs all components of a feature map and apply their weights to produce the output. The output is the probability for each class in a classification task.

4.1.4 ReLu Layer

We can make our input nonlinear by passing the output from the convolution layer to this layer using activation functions. The noise from the convolved feature was eliminated and replaced with a value of 0. Rectified linear units are the optimum answer for situations with a lack of gradient.

4.1.5 Activation Function

Activation functions are non-linear mathematical functions. It estimates non-linear functions, which generates the input. One of the most popular Activation functions is Rectified Linear Unit (ReLU). It can be described as:

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

4.2 Convolution Output Size Calculation

Convolution considers the following parameters:

- Input dimensions: height, width, batch size, and number of channels
- Kernel height and Kernel width
- Stride height and stride width
- Padding height top, Padding height bottom

- Padding width right, Padding width left

We will look at each of the parameters and see how they can be utilized to compute the output size.

4.2.1 Input Dimension

In machine learning models, the input should ideally be four-dimensional (4D), with the following dimensions:

- Height of the input image (H)
- Width of the input image (W)
- Number of batches (N)
- Number of channels (C)

4.2.2 Number of Batches (N)

A batch size (or batch number) refers to how many 2D images are processed together or passed into the machine learning model. The inference process is known as latency when the batch size is one and throughput when the batch size is more than one [55].

4.2.3 Number of Channels (C)

Images are defined by three dimensions: height, width, and channels. The channels determine the features of the photos. It is commonly 3 for RGB or HSV in coloured images. The number of batches (N) and the number of channels (C) are equal for input and output [55].

4.2.4 Kernel

The ideal kernel has two dimensions: height (KH) and width (KW). In addition, it has a third dimension in some cases: channels (KC).

4.2.5 Stride

The value by which the kernel slides over the square segments is called stride. By default, it is 1. Strides can either move along width or height.

4.2.6 Padding

Padding refers to the amount of default data added to the sides of the input to keep the output size consistent. In most cases, the padding area's values are zero. Padding comes in four varieties: on top (P_H1), on bottom (P_H2), on left (P_W1), and on right (P_W2). This modifies the dimensions of the input data as follows [55]:

$$Height = H + P\_H1 + P\_H2 \tag{4.2}$$

$$Width = W + P\_W1 + P\_W2 \tag{4.3}$$

The final formula would be:

$$OutputHeight = (H + P\_H1 + P\_H2 - KH)/(SH + 1) \tag{4.4}$$

$$OutputWidth = (W + P\_W1 + P\_W2 - KW)/(SW + 1) \tag{4.5}$$

A simple example of visualized convolution with 3×3 input, zero padding, a kernel size of 2×2, and a stride of 3 and 2 for height and width respectively is as follows:

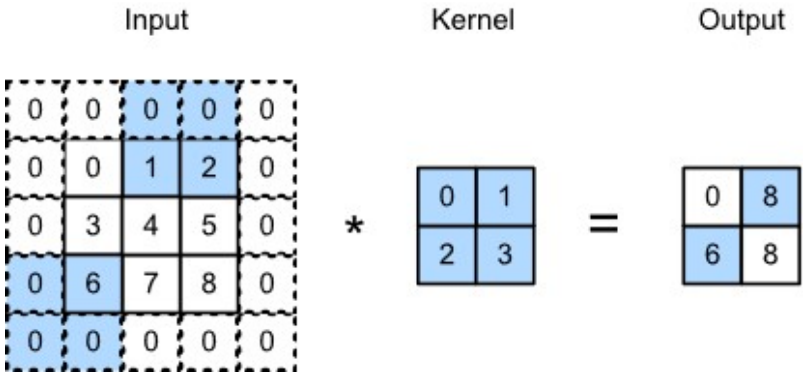

Figure 4.4: An example of visualised convolution with 3×3 input, zero padding, a kernel size of 2×2, and a stride of 3 and 2 for height and width, respectively [8].

4.3 Residual Network (ResNet)

The Fig 4.5 shows that a 56-layer CNN has a greater error rate on both the training and testing datasets than a 20-layer CNN architecture. Normally, overfitting would

result in lower training error. However, the 56-layer CNN also has a higher training error. However, as the number of layers grows, a typical problem in deep learning called "Vanishing/Exploding gradient" emerges.This issue changes the gradient value to either 0 or too large, which prevent the system to learn. As a result, the concept of Residual Network was introduced which works with a smaller number of layers compared with other CNN architectures. ResNet resolves the vanishing/exploding gradient problem by adding the input features to the output. Fig 4.6 demonstrates the residual block in the ResNet architecture.

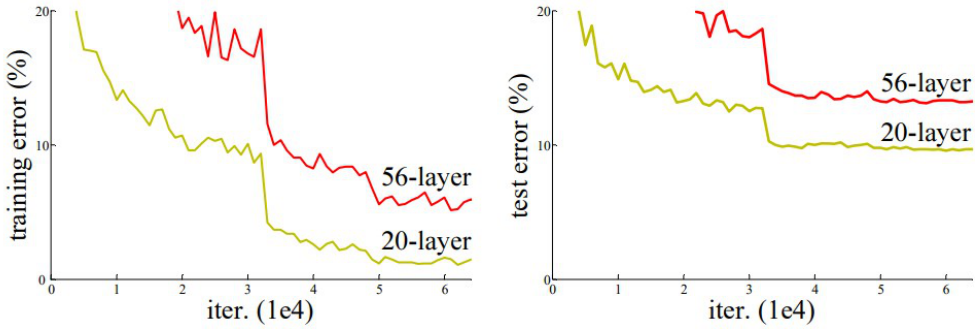

Figure 4.5: Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer "plain" networks. The deeper network has higher training error and thus test error [9].

4.3.1 Residual Block

A method called **skip connections** is employed in this network. The output is linked directly to the training through the skip connection by skipping a few layers to learn. So, instead of say H(x), initial mapping, let the network fit, F(x) := H(x) – x which gives H(x) := F(x) + x.

This skip connection will provide the benefit of bypassing any layers that degrade the performance of the architecture. The results of this enable the training of very deep neural networks without worrying about vanishing/exploding gradients.

4.3.2 Network Architecture

The shortcut connection is implemented after the network is developed based on a 34-layer, plain VGG-19 architecture. By adding ski connection to the design, the design is transformed into a residual network (Fig. 4.7).

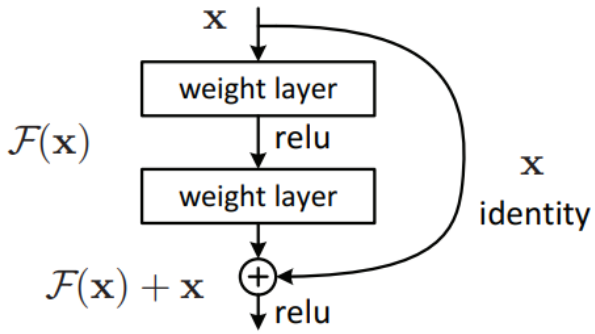

Figure 4.6: Residual learning: a building block [9].

4.4 VGG-16

VGG-16 outperforms AlexNet by replacing large kernel filters sequentially (11 and 5 in the first and second convolutional layers, respectively) with numerous 3x3 kernel filters [56].

4.4.1 Network Architecture

The architecture depicted in Fig 4.8 belongs to VGG-16. Rather than having a large number of parameters, VGG-16 employs 3x3 convolution filter layers with a stride 1. Also, the padding and maximum pooling layer with 2x2 filters and stride 2 remain the same. The convolution and max pool layers are placed in the same way throughout the design. At the end of its architecture, it has two fully-connected layers. The output is then followed by a softmax. The conv1 layer receives a 224 by 224 coloured image as input.

The features are extracted using convolutional layers with the smallest feasible dimensions: 3x3 (to capture left/right, up/down, and centre of images). In one of the VGG variances, an extra 11 convolution filters are added, which may be regarded as a linear change to the input channels (followed by non-linearity). The convolution spatial padding is set to 0 and the convolution stride is set to 1 pixel. After convolution, the spatial resolution of the layer input is preserved, i.e. the padding is 1-pixel for 33% of the convolutional layers. Spatial pooling is done via five max-pooling layers that follow part of the convolutional layers (not all the convolutional layers are followed by max-pooling). Max-pooling is done with stride 2 across a 2x2 pixel frame [57].

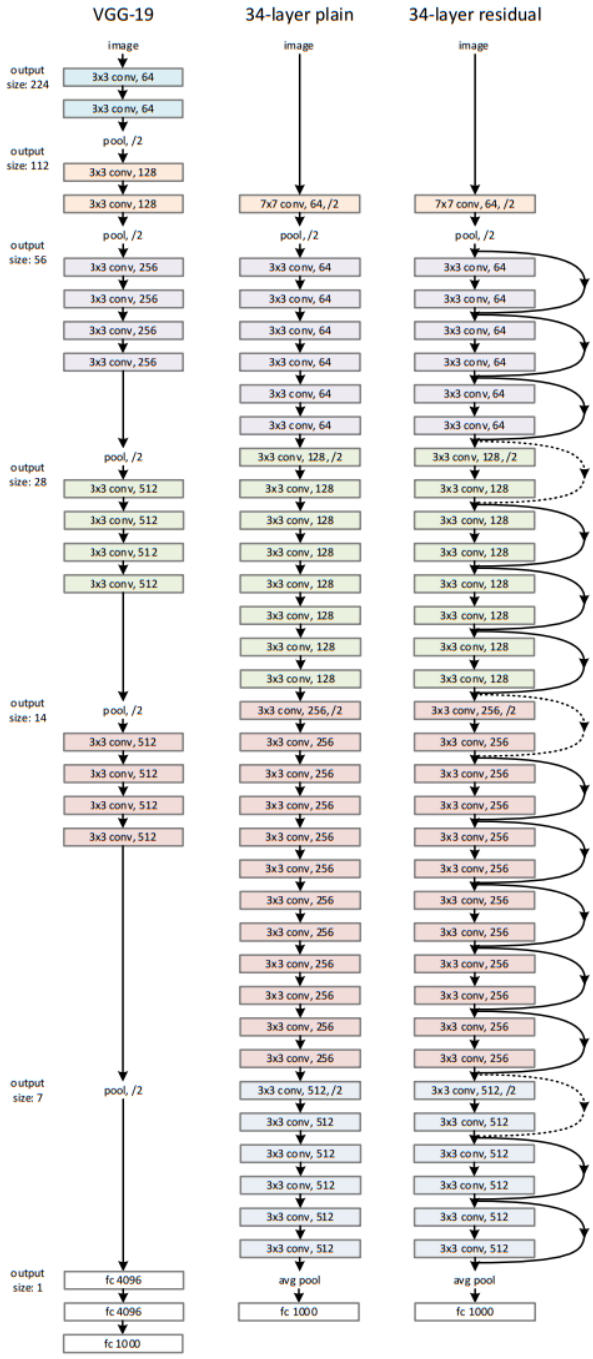

Figure 4.7: Example network architectures for ImageNet. Left: the VGG-19 model [41] (19.6 billion FLOPs) as a reference. Middle: a plain network with 34 parameter layers (3.6 billion FLOPs). Right: a residual network with 34 parameter layers (3.6 billion FLOPs). The dotted shortcuts increase dimensions [9].

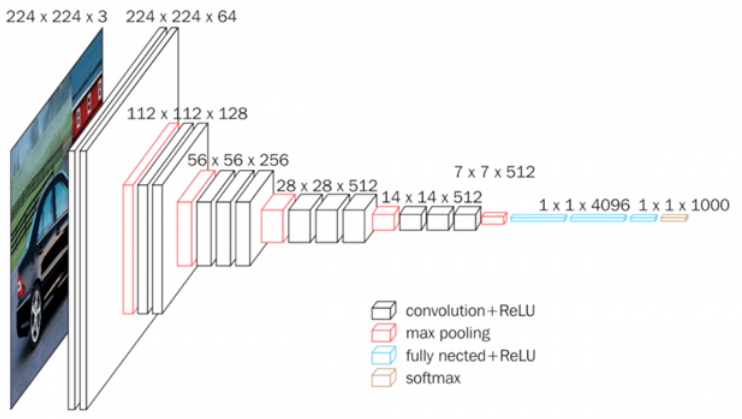

Figure 4.8: VGG-16 network architecture [10].

Following a stack of convolutional layers (of varying depth in various designs), three Fully-Connected (FC) layers are added. 4096 channels are included in the first two FC layers. The last FC layer has 1000 channels since ImageNet dataset contains 1000 classes. The last layer performs as a softmax layer.

Rectification (ReLU) non-linearity is present in all hidden layers. Local Response Normalization (LRN), which does not enhance performance on the ILSVRC dataset but increases memory usage and computation time, is also included in none of the networks. Unfortunately, there are two major drawbacks with VGG Network is its slow speed in to rain, as well as the network architecture weights which are quite large themselves.

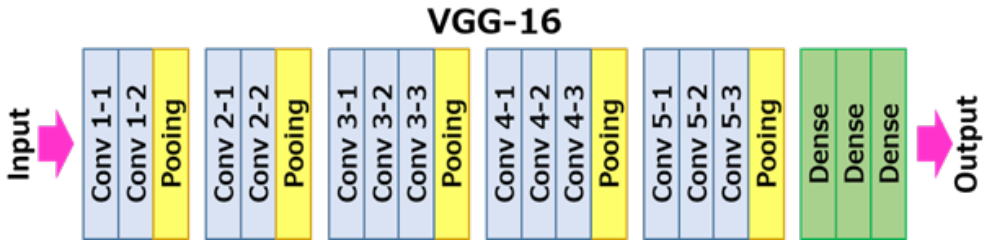

Figure 4.9: VGG-16 network layers [11].

VGG-16 is over 533MB in size because of its depth and amount of completely linked nodes. As a result, installing VGG is a time-consuming process. Many deep learning image classification challenges employ VGG-16; nevertheless, smaller network designs are frequently preferred (SqueezeNet, GoogLeNet, etc.). However, because it is simple to execute, it is an excellent learning tool.

Chapter 5

Evaluation, Analysis and Comparisons

5.1 Dataset

As previously mentioned in Chapter 3, the dataset used in this work is a combination of three different datasets (CK+, EMOTIC, and KDEF), and each of them has its unique features. Firstly, the augmented dataset is trained on a deep neural network, called FaceNet, to extract features from images of a person's face and detect the face. After the face images are detected, they should be augmented by image transformation to be fed to the input of emotion recognition networks. The final dataset consists of cropped, rotated, and horizontally-flipped images of the original dataset. After splitting the total images into training and testing (70% and 30%, respectively), the distribution of seven emotions are shown in Table 5.1, Fig 5.2, and Fig 5.3. As it is shown, the total number of "Happiness" emotion face images outweigh the number of other 6 emotions.

5.2 Evaluation Metrics

5.2.1 Confusion Matrix

Confusion matrix is a performance evaluation metric for machine learning classification problem. The actual target values and predicted values are compared using a confusion matrix. This gives us a holistic view of how well our classification model is

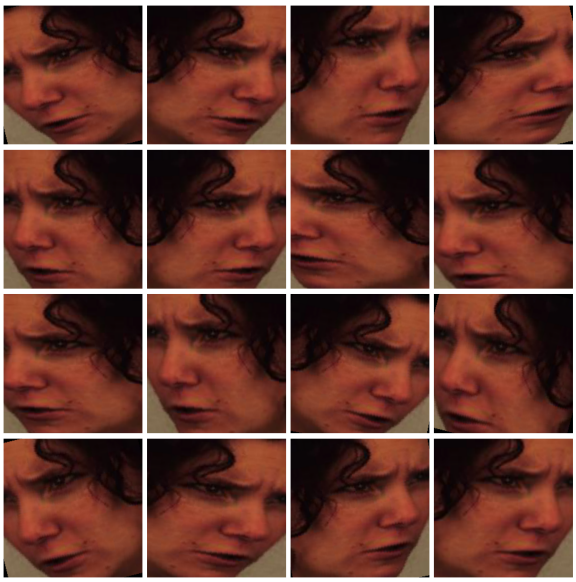

Figure 5.1: An example of standardized input images.

Category	Number of Training Images	Number of Testing Images
Happy	2352	1008
Neutral	1176	504
Disgust	688	295
Sadness	671	287
Surprise	648	278
Angry	617	265
Fear	540	232

Table 5.1: Distribution of training and testing images in seven emotion categories.

Figure 5.2: Distribution of training images in seven emotion categories.

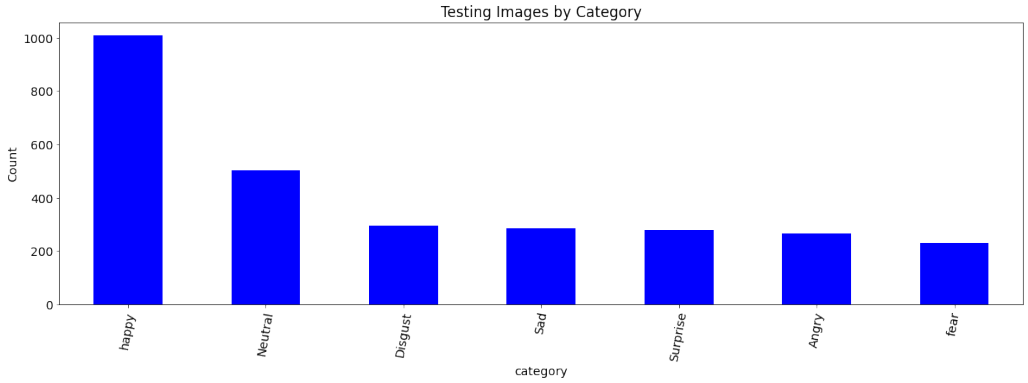

Figure 5.3: Distribution of testing images in seven emotion categories.

performing and what kinds of errors it is making. The target variable has two values: Positive or Negative. The columns represent the actual values of the target variable and the rows represent the predicted values of the target variable (Fig 5.4).

- **True Positive (TP)**: The predicted value matches the actual value. The actual value was positive and the model predicted a positive value.
- **True Negative (TN):** The predicted value matches the actual value. The actual value was negative and the model predicted a negative value.
- **False Negative (FN):** The predicted value was falsely predicted. The actual value was positive but the model predicted a negative value.
- **False Positive (FP):** The predicted value was falsely predicted. The actual value was negative but the model predicted a positive value.

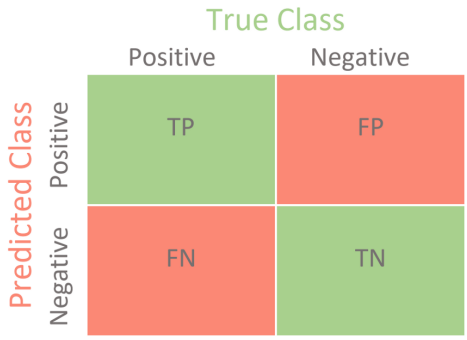

Figure 5.4: Confusion matrix [12].

5.3 Output Result

5.3.1 VGG-16 Results

After running both training and testing datasets on VGG-16 network with 30 epochs, the following results are obtained. Noteworthy, the computational time was about 4 hours.

- **Training Output Results:**
  - **Loss:** It starts from 1.59 at the beginning, and it then plunges to a low of 1.32 after 30 epochs (Fig 5.5).

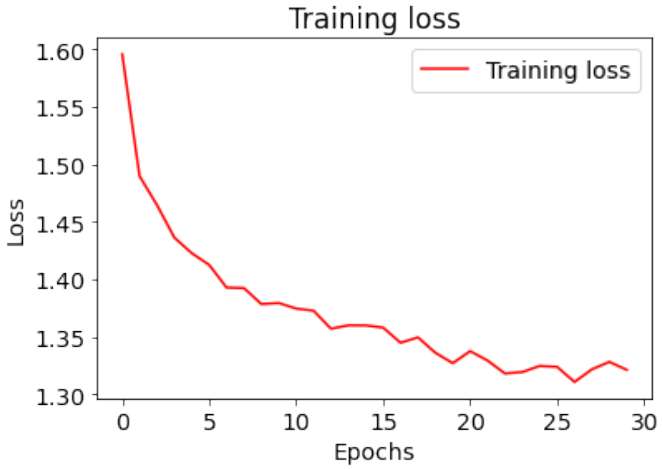

Figure 5.5: VGG-16- Training loss with 30 epochs.

  - **Accuracy:** Staggering soar is observed from epoch 0 with 40.31% accuracy to somewhere in the vicinity of epoch 8 with the value of 50.01%, and it is then followed by fluctuations with a steady increase between epoch 8 and epoch 30 (52.31%) (Fig 5.6).

- **Testing Output Results:**
  - **Loss:** It starts from 1.52 at epoch 1, and it drops to the low of 1.33 after 30 epochs (Fig 5.7).
  - **Accuracy:** It starts with a dramatic increase to epoch 6 and it reaches 51.51% accuracy, and it then fluctuates along with a gradual increase to the end of epoch 30 and it reaches 52.49% accuracy (Fig 5.8).

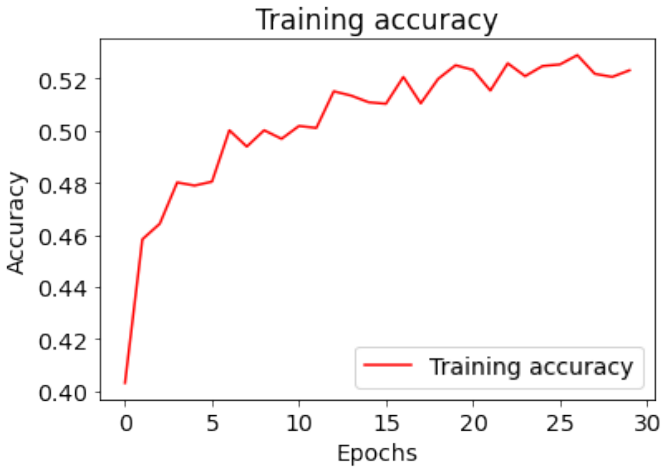

Figure 5.6: VGG-16- Training accuracy with 30 epochs.

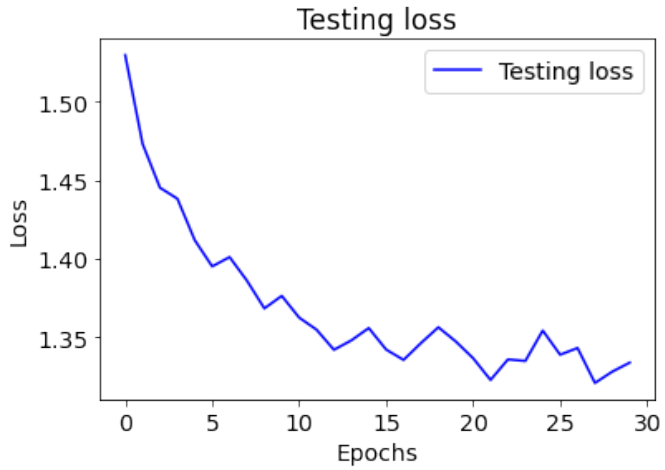

Figure 5.7: VGG-16- Testing loss with 30 epochs.

Moreover, regarding the loss metric, both testing and training dataset decrease exponentially over the graph (Fig 5.9). It is also observed that both testing and training accuracy face the same exponential growth, however, the testing dataset reaches higher accuracy in overall (Fig 5.10).

Through training our dataset on VGG-16 network, the following confusion matrix has been obtained (Fig 5.11). Labels are shown from 0 to 6, which maps to Anger, Disgust, Neutrality, Sadness, Surprise, Fear and Happiness, respectively. The blocks located on the diameter of the matrix show how well each class is distinguished. "Happiness" is the only emotion that reaches high accuracy of determination, while other emotions reach very low accuracy. As it is illustrated,"Happiness" images are signif-

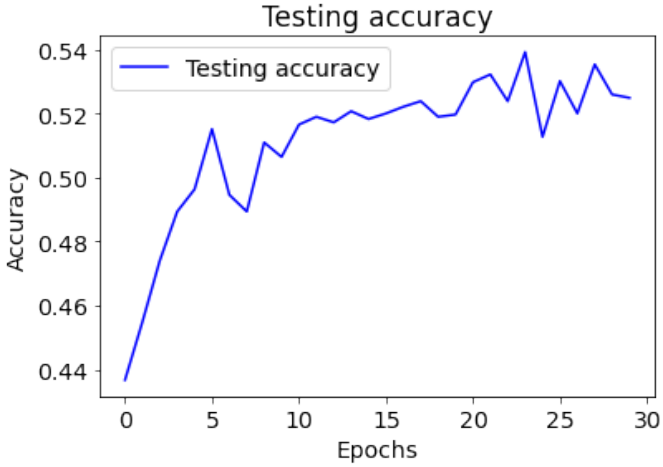

Figure 5.8: VGG-16- Testing accuracy with 30 epochs.

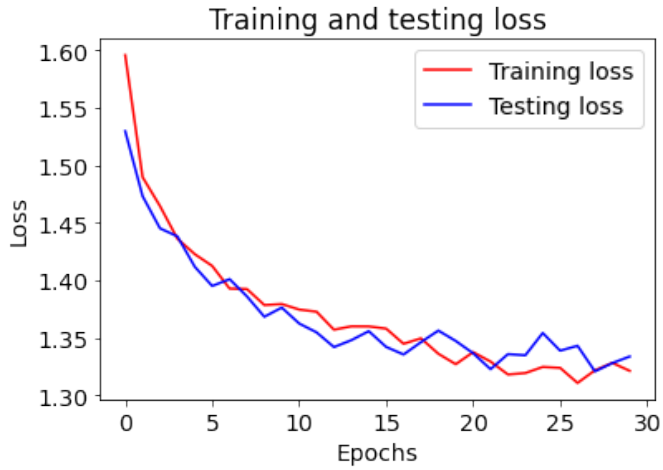

Figure 5.9: VGG-16- Training and testing loss with 30 epochs.

icantly higher in volume compared with other emotion categories. This explains the more accurate prediction of the architecture for this category shown in the confusion matrix.

5.3.2 ResNet50 Results

After running both training and testing datasets on ResNet network with 30 epochs, the following results are obtained. Noteworthy, the computational time was less than 4 hours and it was quite faster than VGG-16.

- **Training Output Results:**

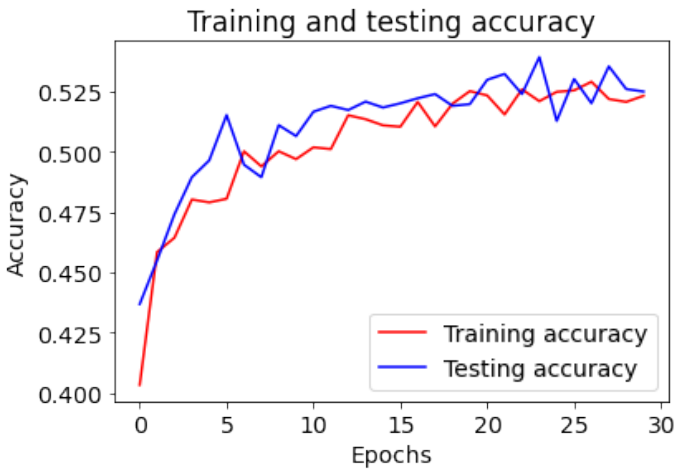

Figure 5.10: VGG-16- Training and testing accuracy with 30 epochs.

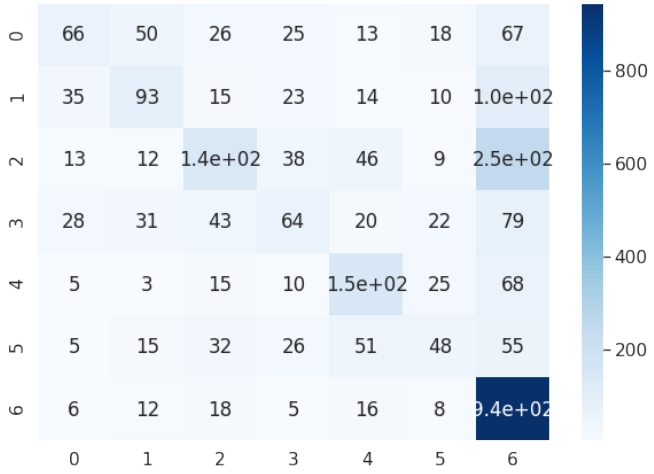

Figure 5.11: VGG-16- Confusion matrix.

- **Loss:** It starts from 1.64 it drops sharply to less than 1.18 after 30 epochs (Fig 5.12).
- **Accuracy:** Considerable soar is observed from epoch 0 with 37.67% to somewhere in the vicinity of epoch 7 with 54.21% accuracy, and it is followed by fluctuations with a steady rise between epoch 8 and epoch 30 (58.39%) (Fig 5.13).

• **Testing Output Results:**

- **Loss:** It starts from 1.55 at epoch 0, and it drops fluctuately to the low of 1.25 after 30 epochs (Fig 5.14).

Figure 5.12: ResNet50- Training loss with 30 epochs.

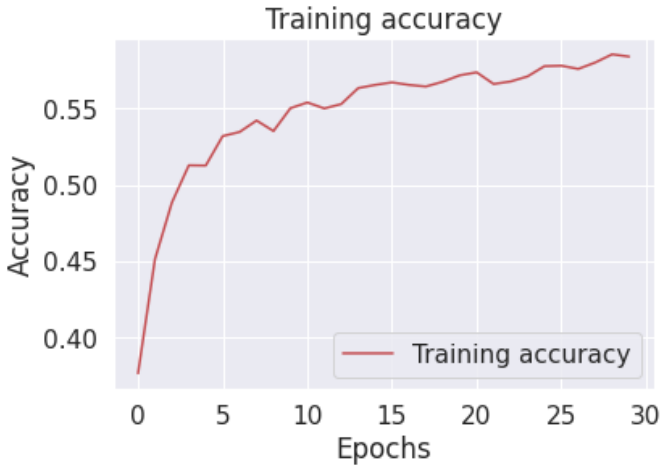

Figure 5.13: ResNet50- Training accuracy with 30 epochs.

– **Accuracy:** It starts from 41.26% with a dramatic increase to epoch 5 and it reaches 52.17% accuracy,s and it then fluctuates along with a gradual increase to the end of epoch 30 with the value of 55.63% accuracy (Fig 5.15).

Moreover, referring to the loss metric, both testing and training dataset decrease exponentially over the graph, while the training dataset reaches lower loss generally (Fig 5.16). It is also observed that both testing and training accuracy face the same exponential growth, however, the training dataset reaches higher accuracy in overall (Fig 5.17).

Through training our dataset on ResNet-50 network, the following confusion ma-

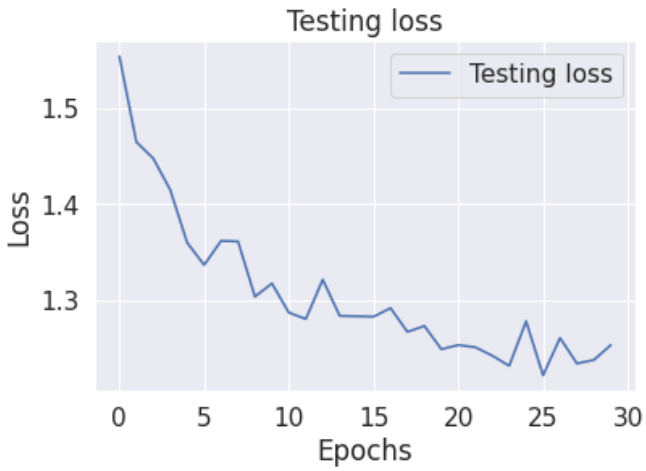

Figure 5.14: ResNet50- Testing loss with 30 epochs.

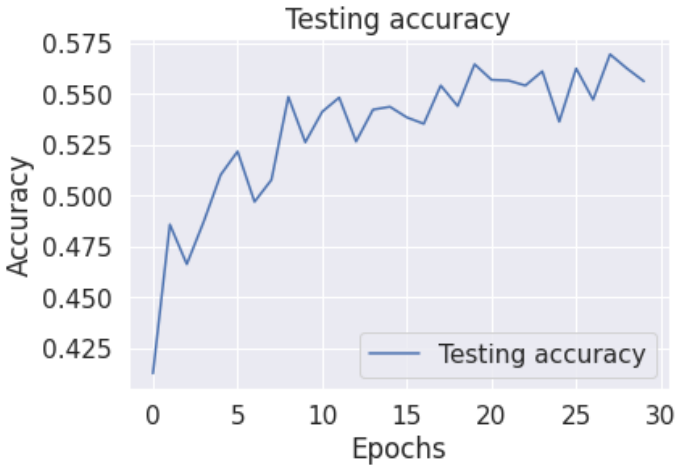

Figure 5.15: ResNet50- Testing accuracy with 30 epochs.

trix has been obtained (Fig 5.18). similarly, "Happiness" is the only emotion that reaches high accuracy of determination, while other emotions reach very low accuracy.

As it is shown in the two above-mentioned confusion matrices, the darker the color of a block is, the higher number of face emotions have been detected. Through analysing the provided confusion matrices, we obtained that the most accurate detection belongs to "Happiness" face images with the number of 940. However, these networks seem to be not effective enough for detecting other 6 emotion in the task of face emotion recognition with the VGG-16 and ResNet networks.

The above confusion matrices show high off diagonal values, which is due to class imbalance. In simpler words, the number of examples in the majority class dominates

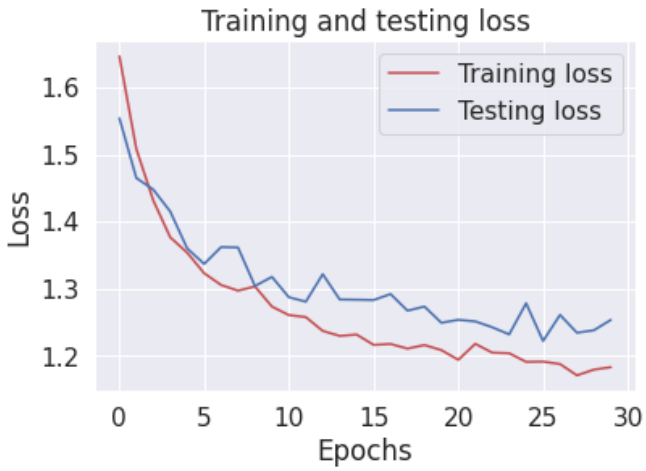

Figure 5.16: ResNet50- Training and testing loss with 30 epochs.

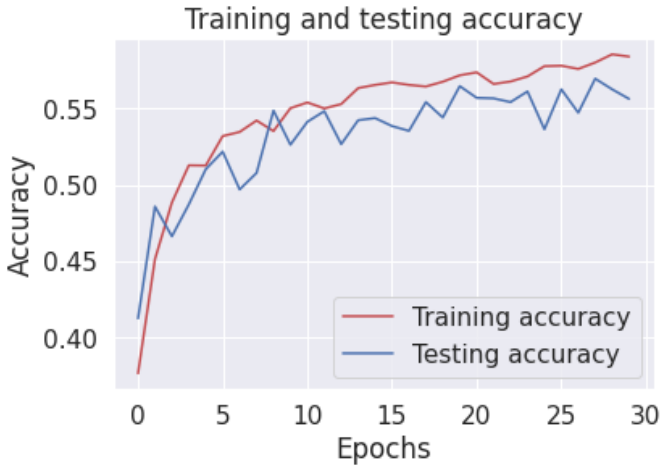

Figure 5.17: ResNet50- Training and testing accuracy with 30 epochs.

the number of samples in the minority class, and the model is unable to learn from minority class well.

There are two approaches to tackle this problem. One approach is to duplicate the images in the minor class. This method is called oversampling and no new information is added to data. Another approach is to randomly delete some data from the major class, which is called undersampling. The main issue with undersampling is a part of data is removed. The approach used in this work is to overcome the imbalanced data using a combination of oversampling and undersampling.

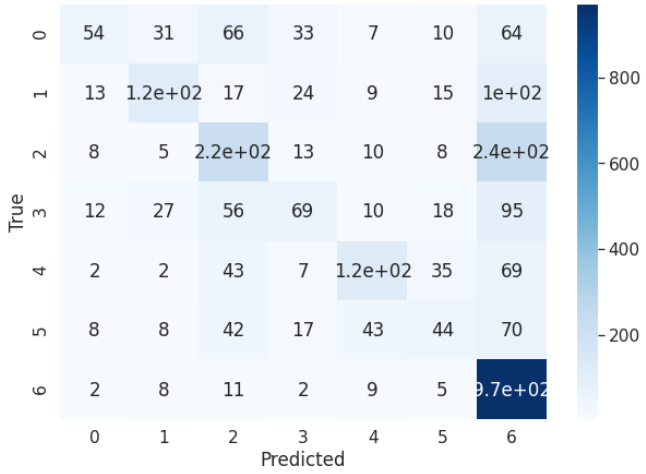

Figure 5.18: ResNet50- Confusion matrix.

5.4 Modification Approach

5.4.1 Synthetic Minority Oversampling Technique

An alternative approach to solve the problem of imbalanced data is to oversample the minority class. This can simply be done by duplicating the examples of minority class, however, this method does not provide any additional helpful information about the data. To defeat this, a method called **"Synthetic Minority Oversampling Technique"**, or **SMOTE** for short is used.

In this approach, the minority class is over-sampled by creating "synthetic" examples instead of replacing the over-sampled examples [58]. The new samples are duplicated based on the Euclidean distance of each data and the minority class nearest neighbors, so the generated examples are different from the original data and provide additional information which is useful for the system to learn the model.

5.4.2 Tomek Link

Another approach to solve the problem of imbalanced data is to undersample the majority class. Previously-mentioned, this can be done by removing random samples from dataset, however, some information are deleted. To defeat this, a method called **"Tomek Link"** is used [59]. Random samples from the majority class are removed in this method. This is also considered as an enhancement of Nearest-Neighbor Rule (NNR) [60]. This method uses the rule to selects the pair of examples that fulfill

specific properties. One of the advantage of this method is it removes the data from the majority class that has the lowest Euclidean distance with the minority class data, therefore make it less ambiguous to detect the emotion.

5.4.3 SMOTETomek

This method is a combination of SMOTE and Tomek link approaches. In other words, it has the ability to oversample by creating synthetic data for minority class as well as the ability to undersample by removing data from the majority class. More Accurate results are obtained by using SMOTETomek for the task of data augmentation. After modifying our dataset, using SMOTETomek approach, improvement over the initial method is observed. These improvements are noticed over both architectures.

5.4.4 VGG-16 Results

After running our modified dataset on VGG-16 network with 30 epochs, the following results are obtained:

- **Training Output Results:**
  - **Loss:** It starts from 1.64 at the beginning and it then plunges to a low of 1.36 after 30 epochs (Fig 5.19).

Figure 5.19: VGG-16- Training loss with 30 epochs on modified dataset.

  - **Accuracy:** Staggering soar is observed from epoch 0 (35.68%) to 47.59% at epoch 30 (Fig 5.20).

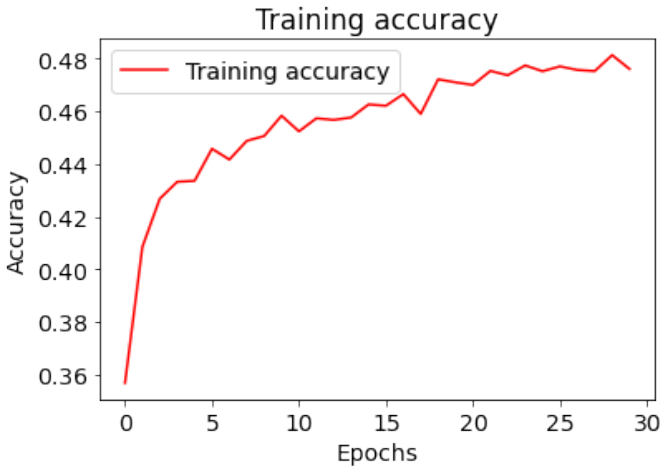

Figure 5.20: VGG-16- Training accuracy with 30 epochs on modified dataset.

- **Testing Output Results:**
  - **Loss:** It starts from 1.50 at epoch 0 and it drops to the low of 1.21 after 30 epochs (5.21) (Fig 5.21

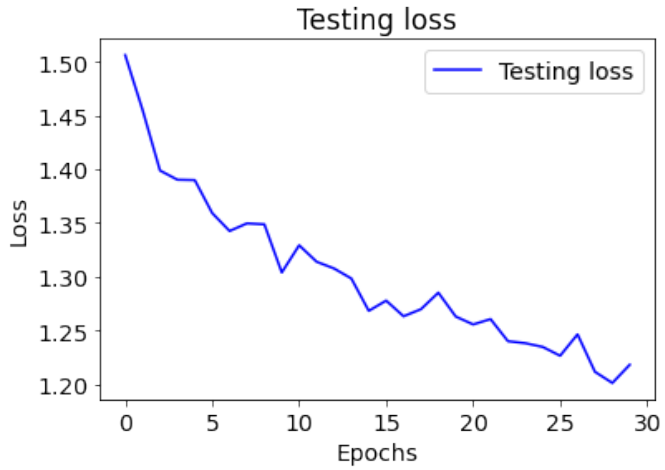

Figure 5.21: VGG-16- Testing loss with 30 epochs on modified dataset.

  - **Accuracy:** It starts form 42.98% accuracy at epoch 0 and it then fluctuates along with a gradual increase to the end of epoch 30 which it reaches around 55% accuracy (Fig 5.22).

Moreover, regarding the loss metric, both testing and training dataset decrease exponentially over the graph, while the testing dataset reaches lower loss generally (Fig

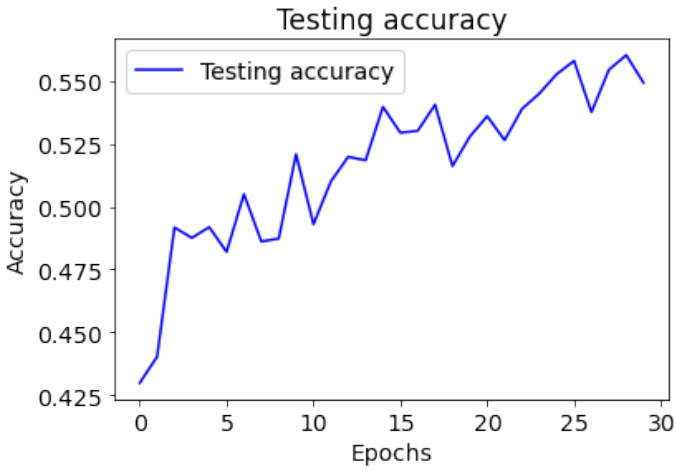

Figure 5.22: VGG-16- Testing accuracy with 30 epochs on modified dataset.

5.23). It is also observed that both testing and training accuracy face the same exponential growth, however, the testing dataset reaches higher accuracy in overall (Fig 5.24). Through training our modified dataset on VGG-16 network, the following confusion matrix has been obtained (Fig 5.25). As it is observed, the blocks on the diameter are darker, which means the classes are classified more accurate when the number of samples in each class is balanced. Noteworthy that higher number of "Happiness" and "Sadness" face images are distinguished correctly, respectively.

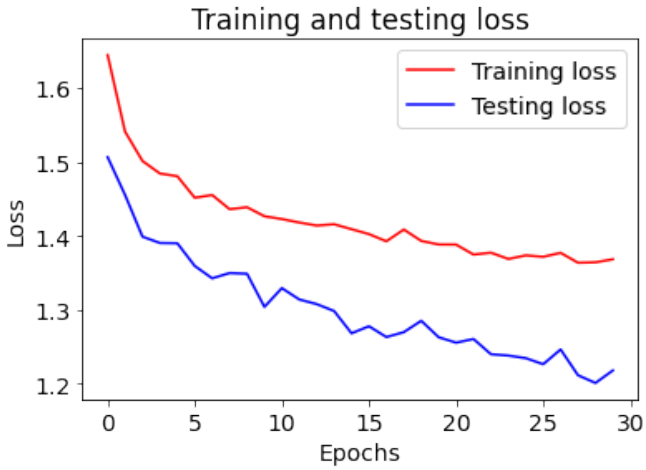

Figure 5.23: VGG-16- Training and testing loss with 30 epochs on modified dataset.

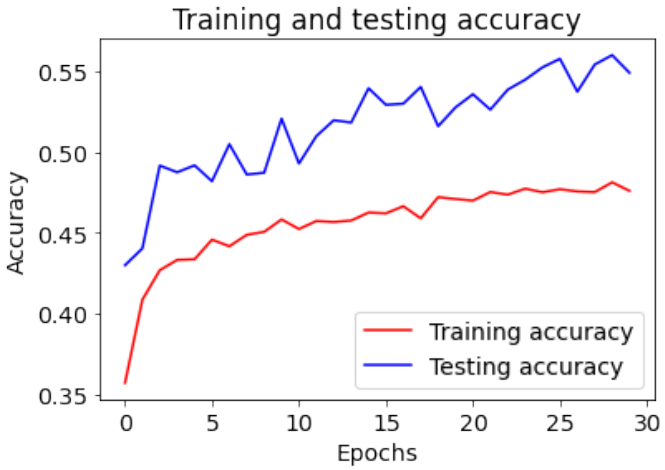

Figure 5.24: VGG-16- Training and testing accuracy with 30 epochs on modified dataset.

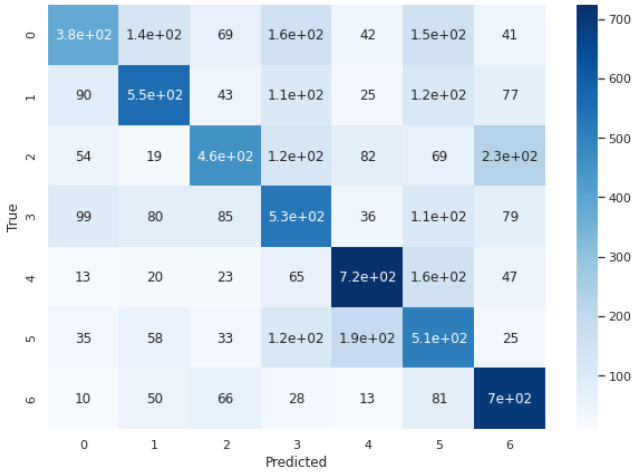

Figure 5.25: VGG-16- Confusion matrix for modified dataset.

5.4.5 ResNet50 Results

After running the modified dataset on ResNet network with 30 epochs, the following results are obtained:

- **Training Output Results:**
  - **Loss:** It starts from 1.66 it drops sharply to around 1.22 after 30 epochs (Fig 5.26).
  - **Accuracy:** Considerable soar is observed from epoch 0 with 34.97% and to somewhere in the vicinity of epoch 10 with 50.06%, and it is then followed

Figure 5.26: ResNet50- Training loss with 30 epochs on modified dataset.

by some fluctuations with a steady rise between epoch 8 and epoch 30 (53.86%) (Fig 5.27).

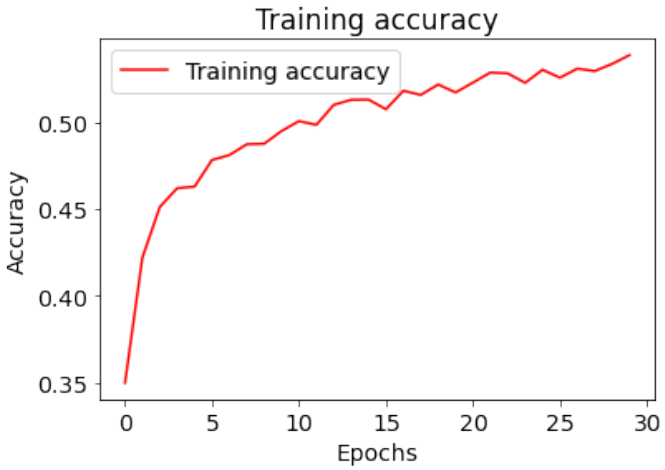

Figure 5.27: ResNet50- Training accuracy with 30 epochs on modified dataset.

- **Testing Output Results:**
  - **Loss:** It starts from 1.52 at epoch 0 and it drops fluctuately to the low of 1.21 after 30 epochs (Fig 5.28).
  - **Accuracy:** It starts with a dramatic increase from 43.59% to 51.90% accuracy at epoch 10 and it then fluctuates along with a gradual increase to the end of epoch 30 which it reaches 53.83% accuracy (Fig 5.29).

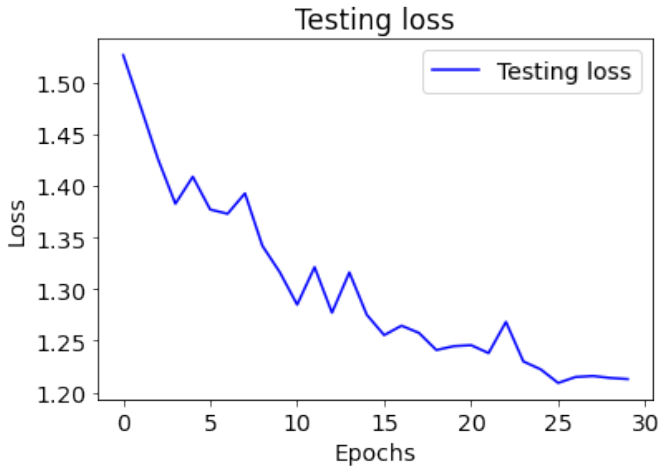

Figure 5.28: ResNet50- Testing loss with 30 epochs on modified dataset.

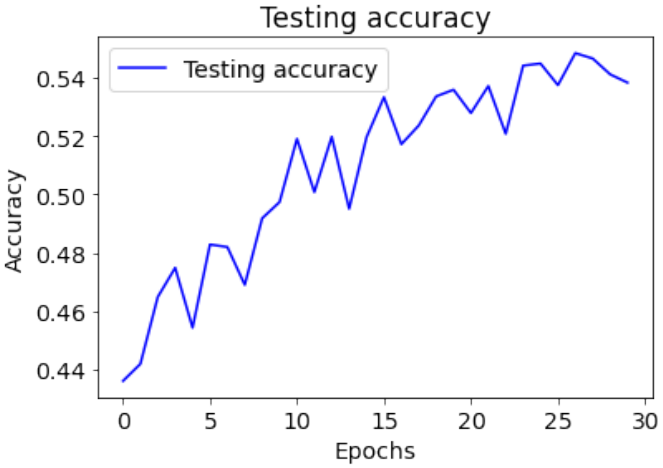

Figure 5.29: ResNet50- Testing accuracy with 30 epochs on modified dataset.

Moreover, referring to the loss metric, both testing and training dataset decrease exponentially over the graph, while the testing dataset reaches lower amount of loss generally (Fig 5.30). It is also observed that both testing and training accuracy face the same exponential growth, however, the testing dataset reaches higher accuracy in overall (Fig 5.31).

Through training our dataset on ResNet-50 network, the following confusion matrix has been obtained (Fig 5.32). Similarly, the blocks on the diameter are darker, which means the classes are classified more accurate when the number of samples in each class is balanced. Noteworthy that higher number of "Happiness" and "Sadness" face images are distinguished correctly, respectively.

The system learns emotion detection well by using the augmented dataset. This is also shown in the confusion matrices obtained (Fig 5.32). As it can be seen the darker colors are dedicated to different blocks of emotion categories. This confirms the effectiveness of the proposed method compared with the initial one. Comparing the enhanced confusion matrices with the initial confusion matrices, more consistency on seven emotion categories is resulted. Still,"Happiness" emotion distinction dominates other emotion detection.

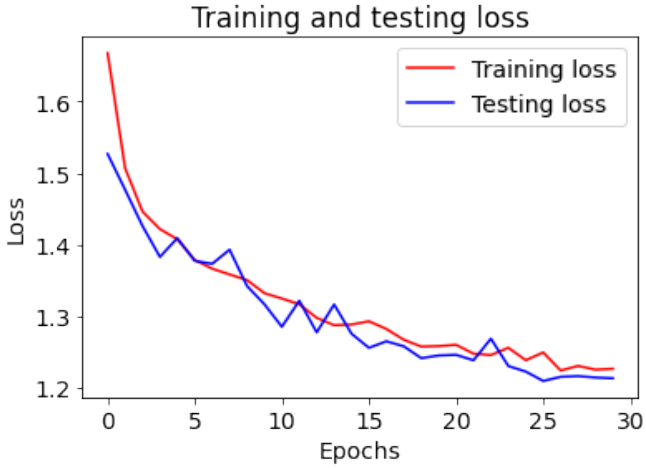

Figure 5.30: ResNet50- Training and testing loss with 30 epochs on modified dataset.

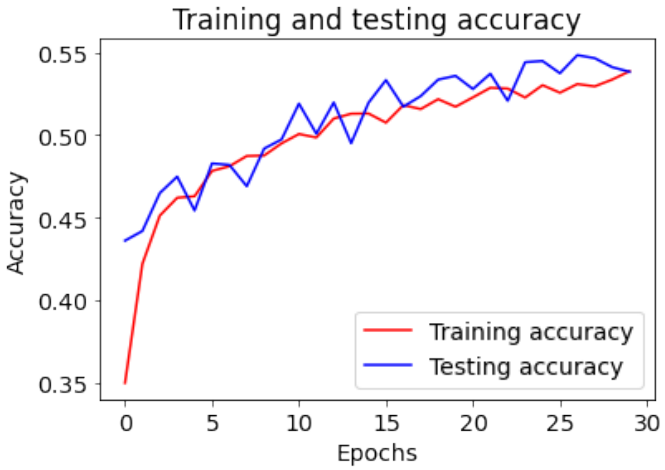

Figure 5.31: ResNet50- Training and testing accuracy with 30 epochs on modified dataset.

The number of images in each category after applying this technique is shown in Fig 5.34. Moreover, the accuracy of initial and enhanced on each network is shown in

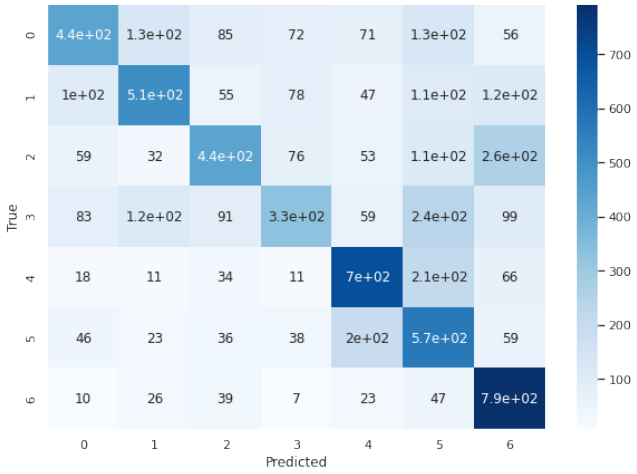

Figure 5.32: ResNet50- Confusion matrix for modified dataset.

	Angry	Disgust	Neutral	Sad	Surprise	Fear	Happy
VGG-Enhanced-Accuracy	0.58301648	0.54151625	0.59459459	0.6116208	0.5625	0.41704442	0.52840159
ResNet-Enhanced-Accuracy	0.58201058	0.59859155	0.56410256	0.53921569	0.60711188	0.40225829	0.54482759
VGG-Initial-Accuracy	0.41772152	0.43055556	0.48442907	0.33507853	0.48387097	0.34285714	0.60295061
ResNet-Initail-Accuracy	0.54545455	0.59701493	0.48351648	0.41818182	0.57692308	0.32592593	0.60323383

Figure 5.33: Initial and enhanced accuracy for each network

	Total Number of Images	Total Number of Images
Emotion	VGG-16	ResNet-50
Angry (1)	789	756
Disgust (2)	1108	852
Neutral (3)	740	780
Sad (4)	327	612
Surprise (5)	1440	1153
Fear (6)	1103	1417
Happy (7)	1514	1450

Figure 5.34: Image distribution in each category after applying the modification approach.

Fig 5.33, which gives a better insight on how well our enhanced dataset has performed. For better comparison, we have shown how VGG-16 and ResNet-50 improved in Fig 5.35 and Fig 5.36. The Y-axis shows the accuracy of the architecture over a certain emotion category. The X-axis represents the emotion categories from 0 to 6 mentioned previously. The enhanced dataset gives priorities to the emotion categories with less data. This improves the accuracy up to 61.16% in VGG-16 and 60.71% in ResNet-50

for certain categories. Also, we can observe consistent behaviour over the enhanced dataset. The summarized results of both networks are addressed in fig. 5.33.

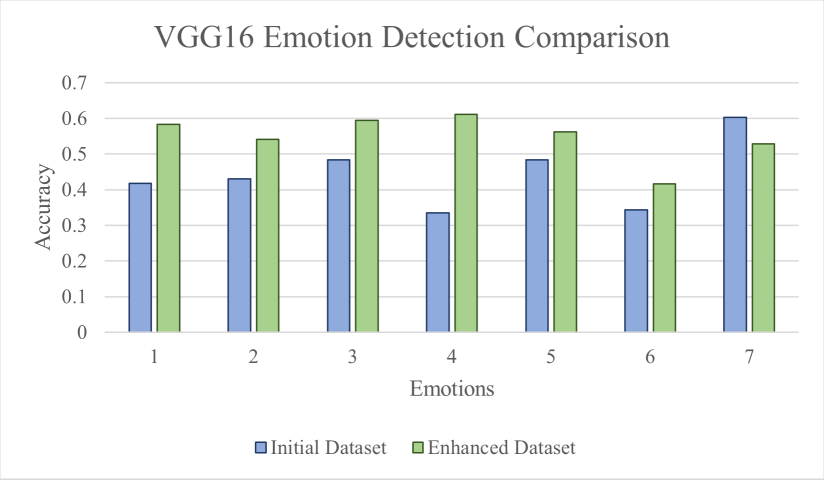

Figure 5.35: Vgg-16 - Emotion detection comparison.

Figure 5.36: ResNet-50 - Emotion detection comparison.

Chapter 6

Conclusions

Because the human face plays such an important part in expressing a person's mental state, facial expression analysis is a major focus of research with numerous potential uses. Scientists from many areas like as psychology, finance, marketing, and engineering have formed a great interest in this subject due to the practical benefits it provides.

This work provides the implementation of facial emotion recognition based on two deep learning algorithms, VGG-16 as well as ResNet-50. From a technical point of view, this work has served to clearly demonstrates the advantage of using a balanced and enhanced dataset including almost same number of examples in each class. The class imbalance data problem is also tackled using a combination of oversampling and undersampling technique, called SMOTETomek. Moreover, the accuracy and loss curves belonging to different steps are provided in the evaluation and analysis sections.

6.1 Future Work

One of the most obvious continuations of this work is the creation of a model that analyse and tackle the reason why performance of some emotions are lower than the other ones. The next step would be to investigate if applying another emotion classification model could be more effective.

Bibliography

[1] Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. EMOTIC: Emotions in Context Dataset. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2017-July:2309–2317, 2017.

[2] Shuxin Ouyang, Timothy Hospedales, Yi-Zhe Song, and Xueming Li. A survey on heterogeneous face recognition: Sketch, infra-red, 3d and low-resolution, 2014.

[3] Contributors. The computer vision pipeline, part 4: feature extraction. https://freecontent.manning.com/the-computer-vision-pipeline-part-4-feature-extraction/, 2019. [Online; accessed 23-August-2021].

[4] PathPartnerTech. Challenges faced by facial recognition system. https://www.pathpartnertech.com/challenges-faced-by-facial-recognition-system/, 2021. [Online; accessed 23-August-2021].

[5] Ellen Goeleven, Rudi De Raedt, Lemke Leyman, and Bruno Verschuere. The Karolinska directed emotional faces: A validation study. *Cognition and Emotion*, 22(6):1094–1118, 2008.

[6] Wikipedia contributors. Neural network. https://en.wikipedia.org/wiki/ Neural_network, 2021.

[7] Medium. The differences between artificial and biological neural networks. https://towardsdatascience.com/the-differences-between-artificial-and-biological-neural-networks-a8b46db828b7, 2018.

[8] Multiple input and multiple output channels colab [mxnet] open the notebook in colab colab [pytorch] open the notebook in colab colab [tensorflow] open the notebook in colab. http://d2l.ai/chapter$_c$*onvolutional* − *neural* − *networks/channels.html*, 2021.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[10] Medium. Convolutional neural network. https://towardsdatascience.com/covolutionalneuralnetworkcb0883dd6529 , 2019.

[11] VGG16 - Convolutional Network for Classification and Detection. Vgg16 - convolutional network for classification and detection. https://neurohive.io/en/popular-networks/vgg16/, 2021.

[12] Medium. Vgg16 - convolutional network for classification and detection. https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826, 2021.

[13] Wikipedia contributors. Emotion recognition — Wikipedia, the free encyclopedia, 2021. [Online; accessed 3-August-2021].

[14] Liudmyla Tereikovska, Ihor Tereikovskyi, Shynar Mussiraliyeva, Gulmaral Akhmed, Aiman Beketova, and Aizhan Sambetbayeva. Recognition of emotions by facial Geometry using a capsule neural network. *International Journal of Civil Engineering and Technology*, 10(03):1424–1434, 2019.

[15] Paul Ekman. FACIAL EXPRESSION Edited by An imprint of The Institute for the Study of Human Knowledge. 1973.

[16] Alex Kelly. Facial expression, 2019.

[17] Human Facial. Book Reviews. (1994):1187–1194, 1996.

[18] Fatemeh Noroozi, Marina Marjanovic, Angelina Njegus, Sergio Escalera, and Gholamreza Anbarjafari. Audio-Visual Emotion Recognition in Video Clips. *IEEE Transactions on Affective Computing*, 10(1):60–75, 2019.

[19] Mohammad Soleymani, Maja Pantic, and Thierry Pun. Multimodal emotion recognition in response to videos. *IEEE Transactions on Affective Computing*, 3(2):211–223, 2012.

[20] Harappa. Types of emotions. https://harappa.education/harappa-diaries/types-of-emotions/, 2020.

[21] Ying Li Tian, Takeo Kanade, and Jeffrey F. Conn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001.

[22] D. Y. Liliana. Emotion recognition from facial expression using deep convolutional neural network. *Journal of Physics: Conference Series*, 1193(1), 2019.

[23] Mark Holzer. Multifractal wave functions on a class of one-dimensional quasicrystals: Exact f() curves and the limit of dilute quasiperiodic impurities. *Physical Review B*, 44(5):2085–2091, 1991.

[24] Qingmei Yao. Multi-Sensory Emotion Recognition with Speech and Facial Expression The University of Southern Mississippi. 2016.

[25] Douglas M. Hawkins. The problem of overfitting. *Journal of Chemical Information and Modeling*, 44(1):1–12, January 2004. Copyright: Copyright 2012 Elsevier B.V., All rights reserved.

[26] Visa Sofa and Ralescu Anca. Issues in mining imbalanced data sets - a review paper. 2005.

[27] Junkai Chen, Zenghai Chen, Z. Chi, and Hong Fu. Facial expression recognition based on facial components detection and hog features. 2014.

[28] Caifeng Shan, Shaogang Gong, and P.W. McOwan. Robust facial expression recognition using local binary patterns. In *IEEE International Conference on Image Processing 2005*, volume 2, pages II–370, 2005.

[29] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 568–573 vol. 2, 2005.

[30] J. Whitehill and C. Omlin. Haar features for facs au recognition. *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 5 pp.–101, 2006.

[31] Shervin Minaee, Mehdi Minaei, and Amirali Abdolrashidi. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors*, 21(9), 2021.

[32] Hui-chuan Chu William and Wei-jen Tsai Min-ju Liao. Facial emotion recognition with transition detection for students with high-functioning autism in adaptive e-learning. *Soft Computing*, (1), 2017.

[33] Shervin Minaee, Amirali Abdolrashidi, and Yao Wang. An Experimental Study of Deep Convolutional Features For Iris Recognition Electrical Engineering Department , New York University , Computer Science and Engineering Department , University of California at Riverside. *In Signal Processing in Medicine and Biology Symposium (SPMB)*, 2016.

[34] Ali Mollahosseini, David Chan, and Mohammad H Mahoor. Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*, 2016.

[35] Hamed Zaer, Ashlesha Deshmukh, Dariusz Orlowski, Wei Fan, Pierre-Hugues Prouvot, Andreas Nørgaard Glud, Morten Bjørn Jensen, Esben Schjødt Worm, Slávka Lukacova, Trine Werenberg Mikkelsen, Lise Moberg Fitting, John R. Adler, M. Bret Schneider, Martin Snejbjerg Jensen, Quanhai Fu, Vinson Go, James Morizio, Jens Christian Hedemann Sørensen, and Albrecht Stroh. An intracortical implantable brain-computer interface for telemetric real-time recording and manipulation of neuronal circuits for closed-loop intervention. *Frontiers in Human Neuroscience*, 15:36, 2021.

[36] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong. Facial expression recognition via a boosted deep belief network. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1805–1812, 2014.

[37] Mihalis A. Nicolaou, Hatice Gunes, and Maja Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, 2011.

[38] Konrad Schindler, Luc Van Gool, and Beatrice de Gelder. Recognizing emotions expressed by body pose: A biologically inspired neural model. *Neural Networks*, 21(9):1238–1246, 2008.

[39] Mohammad Soleymani, Sadjad Asghari-Esfeden, Yun Fu, and Maja Pantic. Analysis of EEG Signals and Facial Expressions for Continuous Emotion Detection. *IEEE Transactions on Affective Computing*, 7(1):17–28, 2016.

[40] Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:1960–1968, 2017.

[41] Leon Gu and Takeo Kanade. *Face Alignment*, pages 291–294. Springer US, Boston, MA, 2009.

[42] Wikipedia contributors. Feature extraction — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Feature$_e$*xtractionoldid* = $1006119523, 2021.[Online; accessed 2 - August - 2021]$.

[43] Peopleqlik Pakistan. Challenges faced by face recognition software in pakistan. https://www.checpos.pk/hr-payroll-software-lahore-karachi-islamabad-pakistan/blog/challenges-faced-by-face-recognition-software-in-pakistan/, 2021.

[44] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June:815–823, 2015.

[45] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June:2892–2900, 2015.

[46] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. DeepFace: Closing the gap to human-level performance in face verification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[47] Margarida V. Garrido and Marília Prada. Kdef-pt: Valence, emotional intensity, familiarity and attractiveness ratings of angry, neutral, and happy faces. *Frontiers in Psychology*, 8:2181, 2017.

[48] G T U Affiliated Colleges, Oracle Academy, Oracle Academy, Oracle Academy, Advanced Computer Science, Information Technology, and Advanced Computer Science. Microsoft COCO. *Eccv*, (June):740–755, 2014.

[49] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic Understanding of Scenes Through the ADE20K Dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.

[50] About kdef. https://kdef.se/home/aboutKDEF.html.

[51] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010*, (July):94–101, 2010.

[52] Paul Ekman Group. Facial action coding system. https://www.paulekman.com/facial-action-coding-system/, 2020.

[53] Aseem Patil and Milind Rane. Convolutional Neural Networks: An Overview and Its Applications in Pattern Recognition. *Smart Innovation, Systems and Technologies*, 195:21–30, 2021.

[54] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, pages 1–10, 2014.

[55] OpenGenus IQ: Computing Expertise Legacy. Calculate output size of convolution. https://iq.opengenus.org/output-size-of-convolution/, 2020.

[56] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[57] VGG16 - Convolutional Network for Classification and Detection. Vgg16 - convolutional network for classification and detection. https://neurohive.io/en/popular-networks/vgg16/, 2021.

[58] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, Jun 2002.

[59] Nguyen Thai-Nghe and L. Schmidt-Thieme. Learning optimal threshold on resampling data to deal with class imbalance. 2010.

[60] A. Elhassan and Al-Mohanna. Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method. 2017.