EMOTION RECOGNITION USING DEEP LEARNING FOCUSING ON THE HAND AND FACIAL EXPRESSIONS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
ANKARA UNIVERSITY

by

Hasanain Jawad RADEEF

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

ANKARA
2024

ÖZET

Yüksek Lisans Tezi

EL VE YÜZ İFADELERİNE ODAKLANAN DERİN OĞRENMEYİ KULLANARAK DUYGU TANIMA

Hasanaın JAWAD RADEEF

Ankara Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Asst. Prof. Dr. Yılmaz AR

Duygusal gösterimlerin yol açtığı davranışsal ve fizyolojik tepkileri kullanarak çeşitli duygusal durumların ölçülmesi, tanımlanması ve tanınması, duygu tanıma olarak bilinir. Tartışma yaratma ve sosyal medya da dahil olmak üzere birçok görevdeki sayısız kullanımı nedeniyle duygu tanımlama çok önemli bir alandır. Bilgisayar sistemleri ve araçlarının insan etkilerini algılayıp yorumlamasını sağlayan bir analiz ve zeka sistemi, akıllı etkileşimi tasarlarken ve sunarken uyumlu bir insan-bilgisayar ekosistemi yaratır.

Son yıllarda, yüz ve el hareketleriyle ifade edilen duyguların tanınmasına yönelik sağlam sistemlerin geliştirilmesine odaklanılarak, duygusal bilgi işlem alanına olan ilginin arttığı görüldü. Çalışmamızda, derin öğrenme tekniklerini kullanarak el ve yüz duygu tanıma alanındaki son gelişmelerin kapsamlı bir incelemesini ve analizini sunuyoruz. Duygu tanıma, insan-bilgisayar etkileşiminde çok önemli bir rol oynuyor ve derin öğrenmeyi kullanarak el ve yüz duygu tanıma alanındaki gelişmelere, zorluklara ve gelecekteki potansiyel eğilimlere ilişkin fikir sağlıyor. Toplum, insan-makine etkileşimini içeren teknolojileri giderek daha fazla benimserken, doğru ve etik açıdan sağlam duygu tanıma sistemleri geliştirmek, kusursuz ve empatik arayüzler oluşturmak için gerekli hale geliyor.

Araştırma ayrıca, iyi bilinen derin öğrenme tekniklerinin karşılaştırmalı bir incelemesini de içeriyor; performans göstergelerini, hesaplama etkinliğini ve dinamik yüz ifadeleriyle karşı karşıya kaldığında dayanıklılığı değerlendiriyor. Ayrıca bu araştırma, modellerin çeşitli alanlara ve veri kümelerine uygulanabilirliğini incelemektedir.

Bu çalışmada, yüz ifadelerinin yedi türe ayrıldığı, yüz hareketi tanıma için yaygın olarak kullanılan iki veri tabanı olan RAF-DB ve FEER 2013 olmak üzere iki veri tabanı kullandık: öfke, tiksinme, şaşkınlık, mutluluk, korku, üzüntü ve nötr. . Bu iki veritabanındaki görüntüleri üç derin tanıma algoritması kullanarak eğittik ve test ettik ve umut verici sonuçlar gösterdik.

Çalışmalarımıza temel olarak geleneksel yöntemle başlıyoruz. Daha sonra 34th ResNet (He vd., 2016), MobileNet-V3(Howard ve diğerleri2017) ve Wider ResNet50-2'nin (Zagoruyko ve Komodakis, 2016) derin öğrenme tekniklerini içeren daha karmaşık bir yönteme geçiyoruz. Görüntünün farklı bölümlerine odaklanıyoruz: bir kanal yüz için,

diğeri sağ el için ve diğeri sol el için. Daha sonra, yüzden duygu tanıma konusunda daha önce yapılan benzer çalışmalardan elde edilen bazı ölçümlere dayalı farklı teknikleri gösteriyoruz.

Yüz İfadesi Tanıma 2013 (FER2013) ve RAdboud Yüzler Veritabanı (RAF-DB) veri setleri üzerinde temel bir model oluşturuyoruz ve çeşitli veriler üzerinde CNN kullanma deneyimimize göre katman başına yeni katman ve nöron sayısını seçiyoruz. Setler. Deneme yanılma yaklaşımı ideal öğrenme oranını ve diğer güçlendirme parametrelerini belirler. ImageNet veri seti üzerinde önceden eğitilmiş ResNet50v2'yi kullanarak el pozisyonuna göre yumruk, süper, kaybeden ve zafer etiketlerini tahmin etme.

%69,41'lik test doğruluğuyla Wider ResNet50-2 (Zagoruyko ve Komodakis, 2016) modeli (Zagoruyko ve Komodakis, 2016), FER2013 veri kümesinde en iyi performansı gösterir. Orijinal Wider ResNet50-2 modeli 68 milyon parametre içermesi nedeniyle diğerlerine göre oldukça karmaşık ve büyüktür. Ancak parametre sayısını 66 milyona düşürdük. ResNet 34 %68,57, Mobile Net V 3 ise %66,09 doğruluk oranına ulaştı.

Daha geniş olan ResNet50-2 (Zagoruyko ve Komodakis, 2016), %87,23 ile RAF-DB'de en iyi test doğruluk puanına ulaşır. Ek olarak, boru hattımızı MTCNN yüz algılama modelini kullanarak uyguluyoruz ve nasıl performans gösterdiğini görmek ve davranışını daha iyi anlamak için gerçek dünya verilerini kullanarak test ediyoruz. Diğer algoritmalarla yaptığımız çalışmalarda kullanılan aynı iki veritabanı için daha düşük doğruluk puanları elde ettik. ResNet 34 ve MobileNet gibi, ResNet 34 %86,32, Mobile Net V 3 ise %83,44 doğruluk oranına ulaştı.

Modelin verimliliğini değerlendirmek için kafa karışıklığı matrisi ve etiket başına hassasiyet, hatırlama ve F1 puanı dahil diğer önlemler kullanıldı. İlk olarak, model tarafından üretilen çok sayıda hatanın yanı sıra çeşitli etiketlerde her hatanın meydana gelme sıklığını gösteren karışıklık matrisine bakalım. Derecelendirme sisteminin etkinliğini arttırmanın bir yoludur. Veri setinde ikiden fazla sınıf varsa veya sınıflar arasındaki gözlem miktarında farklılıklar varsa sınıflandırma doğruluğu zorlayıcı olabilir.

Karışıklık matrisi, doğruluk, hatırlama ve F1-Skorunun hesaplanması yoluyla, belki bir sınıflandırma modeliyle bağlantılı zaferler ve eksiklikler hakkında daha derin bir anlayış kazanabiliriz. Bu faktörler, sınıflandırma sisteminin genel doğruluğundan çok, sistem performansının daha ayrıntılı bir şekilde anlaşılmasını sağlamaya odaklanmıştır.

İyimser tahminlerin doğruluğu doğrulukla ilişkilidir. Geri çağırma ise yalnızca iyi olayların belgelenmesine odaklanır ve F1 puanı hem hatırlamayı hem de kesinliği hesaba katan adil bir değerlendirme sunar. F1 skorunda, harmonik ortalama kullanılarak kesinlik ve hatırlamaya eşit ağırlık verilir.

Ayrıca, optimal sonuçlarımızın RAF-DB ve FER2013 veri setlerinden elde edilen diğer güncel sonuçlarla karşılaştırmalı bir analizini gerçekleştirdik. Bu arama modeli Wider-ResNet50-2'nin doğruluğu, FER2013 ve RAF-DB üzerinde değerlendirildiğinde tüm alternatif arama modellerini geride bırakıyor. Bunun tersine, bazı araştırmacılar birden fazla veri kümesinin tek, daha kapsamlı bir veri kümesinde entegrasyonuna güvenirken,

bizim yaklaşımımız herhangi bir birleştirme metodolojisi kullanmadan her veri kümesini ayrı ayrı kullanmaktır. İlk modellerle karşılaştırıldığında, her model için özelleştirme katmanlarını kullanmamız eş zamanlı olarak çıkarım için gereken süreyi artırır ve toplam model parametresi sayısını azaltır.

Yukarıda belirtilenler aracılığıyla, bu araştırma, derin öğrenme tekniklerini kullanarak yüz ve el hareketlerinden duygu tanıma alanındaki ilerlemelere ve zorluklara kapsamlı bir genel bakış sunmakta ve daha etkileşimli ve güçlendirici deneyimler elde etmek için insan-makine etkileşimi tekniklerinin nasıl geliştirilebileceğinin anlaşılmasına katkıda bulunmaktadır. .

Ayrıca araştırma, insan-bilgisayar etkileşim sistemlerini iyileştirme potansiyeline vurgu yaparak, daha doğru ve bağlama duyarlı bir anlayış için yüz ve el hareketi tanımanın entegrasyonunu araştırıyor. Kaydedilen kayda değer ilerlemeye rağmen.

Ocak 2024, 79 sayfa

Anahtar kelimeler: Duygu Tanıma, Derin Öğrenme, ResNet, Mobile Net

ABSTRACT

Masters Thesis

EMOTION RECOGNITION USING DEEP LEARNING FOCUSING ON THE HAND AND FACIAL EXPRESSIONS

Hasanain Jawad RADEEF

Ankara University
Graduate School of Natural and Applied Siences
Department Of Computer Engineering

Supervisor: Asst. Prof. Dr.Yılmaz AR

Quantifying, describing, and recognizing various emotional states using behavioral and physiological reactions brought on by emotional displays is known as emotion recognition. Due to its numerous uses in many tasks, including discussion generation and social media, emotion identification is a crucial field. A system of analysis and intelligence enabling computer systems and gadgets to perceive and interpret human effects creates a harmonious human-computer ecosystem when designing and presenting intelligent interaction.

In recent years, the field of affective computing has seen an increase in interest, with particular focus on developing robust systems for recognizing emotions expressed through facial and hand gestures. In our work, we provide a comprehensive review and analysis of the state-of-the-art in hand and face emotion recognition using deep learning techniques. Emotion recognition plays a pivotal role in human-computer interaction, providing insight into developments, challenges, and potential future trends in the field of hand and facial emotion recognition using deep learning. As society increasingly embraces technologies that involve human-machine interaction, developing accurate and ethically sound emotion recognition systems becomes essential to creating seamless and empathetic interfaces.

The research additionally incorporates a comparative examination of well-known deep learning techniques, assessing performance indicators, computational effectiveness, and resilience when confronted with dynamic facial expressions. Moreover, this research examines the applicability of models to various domains and datasets.

In this work, we used two databases: RAF-DB and FEER 2013, which are two commonly used databases for facial gesture recognition, where facial expressions are divided into seven types: anger, disgust, surprise, happiness, fear, sadness, and neutral. We trained and tested images on these two databases using three deep recognition algorithms and showed promising results.

In our work, we begin with the conventional method as a foundation. Then we go to a more sophisticated method that incorporates the deep learning techniques of 34th ResNet (He et al., 2016), MobileNet- V3(Howard et al., 2017), and Wider ResNet50-2 (Zagoruyko and Komodakis, 2016). We focus on different parts of the image: one channel for the face, the other for the right hand, and the other for the left hand. Then, we demonstrate different techniques based on some metrics from similar previous work in facial emotion recognition.

We create a baseline model on the Facial Expression Recognition 2013 (FER2013) and RAdboud Faces Database (RAF-DB) data sets, and we select the new number of layers and neurons per layer according to our experience in using CNN on a variety of data sets. The trial-and-error approach determines the ideal learning rate and other augmentation parameters. Predicting punch, super, loser, and victory labels based on hand position using ResNet50v2, pre-trained on the ImageNet data set.

With a test accuracy of 69.41%, the Wider ResNet50-2 (Zagoruyko and Komodakis, 2016) model (Zagoruyko and Komodakis, 2016) performs best on the FER2013 dataset. The original Wider ResNet50-2 model is very complex and large compared to others because it includes 68 million parameters. However, we reduced the number of parameters to 66 million, While the ResNet 34 achieved an accuracy rate of 68.57% and the Mobile Net V 3 had a ratio of 66.09%.

The wider ResNet50-2 (Zagoruyko and Komodakis, 2016) achieves the best test accuracy score on RAF-DB with 87.23%. Additionally, we implement our pipeline using the MTCNN face detection model and test it using real-world data to see how it performs and to understand its behaviour better .

The confusion matrix and other measures, including per-label precision, recall, and F1 score, were used to assess the model's efficiency. First, let us look at the confusion matrix, which illustrates the multitude of errors generated by the model as well as the frequency at which each error occurs along the various labels. It's a means of enhancing the rating system's efficacy. If there are more than two classes in the data set or if there are differences in the number of observations across classes, classification accuracy may be challenging.

Through the computation of the confusion matrix, accuracy, recall, and F1-Score, we might perhaps gain a deeper comprehension of the triumphs and shortcomings linked to a classification model. These factors are more focused on creating a more detailed understanding of the system's performance than they are on the classification system's overall accuracy.

The accuracy of optimistic predictions is correlated with accuracy. Recall, on the other hand, focuses on documenting only good occurrences, and the F1 score offers a fair assessment that accounts for both recall and precision. In the F1 score, precision and recall are given equal weight using the harmonic mean.

Furthermore, we conducted a comparative analysis of our optimal outcomes with other recent results obtained from the RAF-DB and FER2013 datasets. The accuracy of this search model, Wider-ResNet50-2, surpasses that of all alternative search models when evaluated on FER2013 and RAF-DB. Conversely, while certain researchers depend on the integration of multiple datasets into a single, more extensive dataset, our approach is to utilize each dataset in isolation, without employing any merging methodology. Compared to the initial models, our utilization of customization layers for each model simultaneously increases the time required for inference and decreases the total number of model parameters.

Through the above, this research provides a comprehensive overview of the progress and challenges in the field of emotion recognition from facial and hand movements using deep learning techniques and contributes to understanding how to improve human-machine interaction techniques to achieve more interactive and empowering experiences.

Furthermore, the research explores the integration of facial and hand gesture recognition for more accurate and context-aware understanding, with an emphasis on the potential for improving human-computer interaction systems. Despite the remarkable progress that has been achieved.

January 2024, 79 pages

Key Words: Emotion Recognition, Deep Learning, ResNet, Mobile Net.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to everyone who has contributed to successfully completing my graduation research.

First and foremost, I am deeply thankful to my advisor(Asst. Prof. Dr.Yılmaz AR) for their guidance, support, and invaluable insights throughout the research process. Their expertise and encouragement played a crucial role in shaping the direction of this study.

I am also indebted to the faculty members of the (DEPARTMENT OF COMPUTER ENGINEERING) at (ANKARA UNIVERSITY) for their constructive feedback and scholarly input. Their commitment to academic excellence has been a constant source of inspiration.

Special thanks go to my classmates and friends who provided moral support and engaging discussions, enriching the research experience.

Finally, I extend my heartfelt appreciation to my family for their unwavering support and understanding during the ups and downs of this academic journey. Their love and encouragement have been my driving force.

This research would not have been possible without the collective efforts of all those mentioned above. I am truly grateful for the opportunity to undertake this study and the invaluable lessons learned along the way.

Hasanain Jawad RADEEF
Ankara, January 2024

TABLEOF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

The topic's relevance is evident in Shakespeare's "Macbeth" (1699), where King Duncan faces betrayal from Thane of Cawdor. It turns out that measuring someone's thoughts or trustworthiness is impossible based on their appearance alone.

This quote remains applicable in situations today, particularly regarding misunderstandings in communication. However, it is undeniable that we have developed an ability to interpret emotions through expressions. This work aims to explore the extent to which facial cues reliably convey emotions and whether this mode of communication is equally effective in perceiving someone's state.

Gestures are a type of nonverbal communication. By utilizing gestures involving the hands, head, and several other body parts, A variety of ideas, feelings, and emotions can be expressed by people (Allen & Paese 2004; Ekman et al. 1987; Oster 1988). According to Pease and Pease (2008), gestures can be classified into three groups;

- Nodding is likely a way for people to show affirmation or agreement; as a result, even those born blind utilize it as a form of communication. This nodding is an example of an intrinsic gesture.
- Extrinsic: Turning one's body to the side as a kind of denial is an example of the extrinsic type. This extrinsic is a sign that we learn to make when we are very young, for occasions such as when a newborn has consumed enough breast milk from their mother.
- Because of of natural selection, one possible illustration of this would be the widening of the nostrils to increase the amount of oxygen that enters the body; this may be interpreted as a sign of preparation for either combat or flight.

Computers are now an integral part of every person's existence; thus, having an understanding of the human emotional state would enable the computer to adapt more effectively and would, in general, increase collaboration. The ability to recognize

gestures is becoming an increasingly significant field of application for computer vision. Gesture is an essential mode of communication (Rosenstein and Oster 1988).

In the context of our issue, our goal is to construct the system in stages, progressing from an easy method to a difficult one and from straightforward information to an involved one. In order to go further, we want to adopt a method of learning that is gradual.

Further investigation into the issue and an in-depth understanding of how computer vision may be used for more complex and complicated activities is required.

Body language is a broad term encompassing various forms of non-verbal communication, such as facial expressions, body positioning, hand gestures, eye movements, and physical contact, physical contact, and personal space usage. By observing someone's posture, the position of their hands and legs, and how they carry themselves while standing, sitting or moving around, we can deduce their emotional state and overall wellbeing.

In this study, the faces and hands of humans are singled out as the two human body parts that are most crucial to comprehending the meaning of a person's gestural expressions. After the face, the information provided by a person's hands is perhaps the most abundant in their body language. In addition, the position of the head and how it moves may be quite informative; for instance, the speed with which someone nods might indicate their level of patience. Exposing the neck may be seen as a gesture of surrender.

Because of the considerable degree to which culture influences gestures, cultural differences also play a significant part in our difficulty (Efron 1941, Kendon 1983). The ideal option is to globalize some gestures since all humans share specific fundamental movements. Each complicated emotion may be a combination of these many worldwide

gestures. The following table provides a concise illustration of these fundamental hand motions.

Since men and women communicate via body language in fundamentally different ways, gender plays a significant influence in this area, and the same gesture may convey an entirely different set of emotions depending on who is gesturing (Allen& Paese 2004). Men have less facial expression than women; they smile less and exhibit less emotion. This contrast in facial expression is true across the board. Women are expected to get along with people and collaborate constantly, and they are also taught greater appeasing body language from the time they are infants (Gunes and Piccardi 2005).

Gender plays a significant role in this context because of the noticeable discrepancies in how individuals of different genders express themselves through body language. Additionally, the same gesture can convey varying emotions depending on the gender of the person doing it. The source cited is Allen and Paese's publication from 2004. In general, males have fewer expressive facial expressions than women do; apart from smiling, they do not show any emotion in their faces. This facial expression is because women are trained to have more appeasing body language as youngsters and are constantly pressured to cooperate and get along with others. Another reason is that women are encouraged to always get along with people and work together (Allen & Paese 2004).

In the course of our work, we begin with the conventional method as a foundation, and then we go to a more sophisticated method that incorporates deep learning techniques of 34th ResNet (He et al., 2016), MobileNetV3(Howard et al., 2017), and Wider ResNet50-2 (Zagoruyko and Komodakis 2016).

We focus on different aspects of the image: one channel for the face, the other for the right hand, and the other for the left hand. Then, we will compare techniques based on metrics similar to previous work in Facial Emotion Recognition.

Figure 1.1 Numerous nonverbal cues are included in body language, including touch, head and hand positions, eye movements, body posture, facial emotions, and gestures. (Gunes and Piccardi 2005).

Table 1.1 Protocols for the six basic emotions (Gunes and Piccard 2005, 2006, Gunes et al., 2015)

Emotion	Associated Gestures
Fear	Elevated heart rate. Crossing limbs and making movements with arms and legs. Wrapping legs around objects, engaging in bouncing motions, gripping with hands or arms, pulling elbows inward. Tense muscles, holding breath, maintaining a posture. Heightened alertness reflected in body language.
Anger	Stand with your arms relaxed and slightly apart. Place your hands on your hips or waist. Keep your fists either clenched or closed. Ensure that your palms are facing downwards. When necessary, Raise the right hand. Use one hand to point your finger. Avoid shaking hands. Feel free to cross your arms when comfortable.
Sadness	A person collapsed. Their body hunched forward with their hands covering their face. They seemed surprised as their arms wrapped around themselves. They shrugged their shoulders. Their hands were positioned lower than usual, either closed or moving in a manner. Gradually, they. Moved their hands over their head. At one point, they placed a hand on their neck. Clenched both hands together. Their head was tilted to the side.

Table 1.1 Protocols for the six basic emotions (Gunes and Piccard 2005, 2006, Gunes et al., 2015) (continue)

Emotion	Associated Gestures
Surprise	Suddenly, travelling in the other direction. Bringing one or both hands up to the head in a similar manner. Extending one's hand. Bringing both hands to the head and touching it when the mouth is touched. Hands-on the face, either one or both. Bringing both hands up to the level of the head. Put a hand on the face and look at it. Oneself. Using both hands to cover the lips or cheeks of the individual. Shaking the head occurs after stepping back or altering one's body posture.
Happiness	I extend my arms and swing them. I stretch out my legs, keeping them straight. I might spread my legs apart with my feet directed towards something or someone of importance. As I look around, I relax my eye contact. Allow it to linger for a while.
Disgust	Expressing feelings of disgust and support, individuals tend to protect their necks by placing their hands around them. They may also cover their mouth with one hand. Bring the hand close to their body. In certain situations, people might shift their bodies to one side. Change their orientation while ensuring that they keep their head covered with their hands.

2. LITERATURE REVIEW

2.1 Introduction

Over the years, researchers have extensively used indicators of emotion such as facial expressions, action units, and valence arousal to study emotion identification on a large scale. These investigations have spanned decades.

Convolutional neural networks (CNNs) have performed well in image processing since their introduction in the 1990s because of their adaptability and variety (LeCun, 1998). The convolutional, pooling, and fully connected layers are utilized in convolutional neural networks (CNNs). Convolutional Neural Networks (CNNs), often composed of multiple layers, enhance images. However, using CNNs was greatly hindered during that specific time frame due to the limited training data and the low computing capabilities. During the 2010s, improvements in computer power and the availability of bigger datasets made Convolutional Neural Networks (CNNs) a more viable method for extracting features and classifying images (Heravi et al., 2016).

Deep learning algorithms that use networks have grown more popular in recent years to assess visual data, such as movies and television pictures. The trajectory that is now being followed is anticipated to continue upward. It has been shown via empirical research that convolutional neural networks (CNNs) such as ResNet, VGGNet, and AlexNet are effective in picture categorization and feature extraction. These networks were established by AlexNet and VGGNet, who are responsible for their establishment. The notion of the short-term memory network (LSTM), an extension of neural networks (RNNs), was initially introduced by Hochreiter and Schmidhuber in 1997. This short-term memory network is an additional advance in the field. The Long Short-Term Memory (LSTM) model is a crucial tool for natural language processing and video analysis because of its ability to acquire and store information efficiently. 1997 is the year that Hochreiter and Schmidhuber are recognized as being the ones who developed this network.

Several experiments have been conducted recently that have tried to concurrently master these three facial behaviour tasks by using all emotion representations. These studies have been conducted in various contexts, including, but not limited to, clinical settings, observational settings, and laboratory settings. The research undertaken by Kollias and colleagues (Kollias et al., 2019) was the first study to investigate the total integration of all facial behaviour tasks inside a single framework. They came up with the concept of the Face Behaviour Net as a direct consequence of this information. In order to link the training activities, they provided two distinct ways. They used several distinct emotion databases that were easily accessible and did not need anything to be paid.

Kollias and his colleagues introduced the Aff Wild2 dataset, which eventually developed into a comprehensive database that included annotations on all three core behaviour tasks (Kollias and Zafeiriou 2019). Furthermore, they developed multitasking learning models that used visual modalities and the ArcFace loss function for emotion detection (Deng et al., 2019). These models were designed to identify emotions.

A further experiment was conducted by Kollias and his colleagues (Kollias et al., 2021, Kollias and Zafeiriou 2021) with the express objective of addressing the issue of overlapping annotations in datasets used for multitasking learning. Their investigation looked into the extent to which different tasks were correlated with one another. A distribution matching strategy was proposed, allowing information sharing across different tasks by aligning the prediction distributions of those tasks for each other. In order to implement this strategy, the probability distributions of their predictions were aligned. Additionally, Notable occurrences, such as the inaugural Affective Behaviour Analyse in the Wild (ABAW) Competition (Kollias et al., 2020) and the 2020 IEEE Face and Gesture Recognition Conference, were placed concurrently at the same location as the annual conference. Both of these events were held in 2020. With the Aff Wild2 dataset as the foundation, these events significantly created cutting-edge algorithms for evaluating aspects, classifying behaviours, and determining action units.

2.2 Datasets

Here, a compilation of open-source datasets will showcase that are often used in the domain of emotion recognition. The datasets include several characteristics, namely the seven emotions: anger, contempt, fear, happiness, sorrow, surprise, and neutrality. In addition, several datasets specifically target valence and arousal values, which represent the emotional state on a scale ranging from (1) to (1). Specific information on these datasets may be seen in Fig 2.1.

Emotional arousal refers to the physiological and psychological state of heightened activity in response to emotional stimuli. It is closely linked to the "fight or flight" response, a natural reaction to perceived threats or stressful situations.

On the other hand, valence describes the pleasantness or unpleasantness associated with a stimulus. This dimension allows us to classify events and experiences, such as expressions, sounds, music, art forms, pictures, and written or spoken language, along a continuum of positivity or negativity.

FER2013 48 x 48-pixel monochrome images of faces are included in the FER2013 collection. In every image, faces are automatically aligned to occupy the same space and be roughly cantered. This alignment ensures that the facial features stay the same in every image.

This activity aims to classify each face according to its emotional expression into one of seven distinct categories: 0 indicates anger, 1 indicates disgust, 2 indicates fear, 3 indicates happiness, 4 indicates sadness, 5 indicates surprise, and 6 indicates neutrality. The training set has (28,709) occurrences, while the testing set comprises (3,589) instances.

Aff-Wild2: A total of (431) participants were annotated by Aff-Wild2 over (539) movies with a combined total of (2,595,572) frames. Of these subjects, (265) were male

and (166) were female. In a manner that is not reliant on the subjects themselves, the data set has been segmented into the train, validation, and test parts, respectively, including (253, 71), and (233) participants.

AffectNet has amassed a collection of over 440,000 face pictures from search engines that have been personally annotated. During the training phase, we only utilized photographs that included neutral and six essential emotions, which brought the total number of images to roughly (280,000).

RAF-DB The Real-world Affective Faces Database has about 30,000 photos depicting individuals' facial expressions, each of which has been annotated with a basic or complicated emotion (**RAFDB**). Only the 12,271 responses labelled with primary feelings were utilized for the training component of the study.

The RECOLA dataset, created by Ringeval et al., indeed focuses on emotions. However, it primarily targets the analysis of spontaneous emotions in the context of remote collaborative work. Electrodermal activity, electrocardiogram, auditory, and visual are the four modalities in the corpus. It consists of 46 different topics documented in French—around 9.5 hours' worth of recordings. Annotations were included in the audio files. Three female and one male annotators fluent in French each get five minutes to contribute. The dataset is divided into three sections: validation (15 subjects), training (16 individuals), and test (15 participants). Each class has a balanced distribution of pupils based on gender, age, and mother tongue.

AFEW Several This dataset comprises dynamic and temporal facial expressions captured from real-life situations depicted in movies and reality television programs. These expressions have been taken from movies and reality television programs. There are a total of 1809 videos from which selection can made. The whole of the data set may be partitioned into three separate groups; specifically, the training set contains (773 videos), the validation set contains (383 videos), and the test set contains (773 videos) (653 videos). 114 of the 653 video clips in the test set are genuine TV footage since this knowledge will make the assignment more challenging.

On the other hand, most of the training set consists of actual footage and the validation set consists of actual film footage. There are over 330 persons taking part in this event, and the ages of those taking part vary from one to seventy-seven. When generating annotations, one must consider the several expressions that may be seen on a person's face, including wrath, disdain, fear, happiness, neutrality, sadness, and surprised). The Emotion tasks categorize each clip's audio-visual content into the seven fundamental emotional groups.

AFEW-VA Recently, The AFEW-VA database (Kossaifi et al., 2017) was created by annotating a portion of the AFEW dataset, which was acquired via the EmotiW tasks, with valence and arousal labels. The dataset comprises 600 video clips sourced from feature films. These clips accurately replicate real-life settings, including occlusions, varying lighting conditions, and movements without a specific subject. These video excerpts were extracted from full-length movies. The films range in length from brief (less than ten frames) to extensive (over 100 frames) (over 120 frames). This dataset contains annotations about valence and arousal categorized on a frame-by-frame basis. Almost 30,000 frames were assigned specific values within the range of [10, +10] to anticipate the dimensional influence of arousal and valence.

Table 2.1 Gesture Recognition Databases

Database	Labels	Num.of Frames	Num.of Videos
AFEW-VA (Kossaifi et al., 2017)	valence-arousal (discrete)	30,050	600
RECOLA (Ringeval et al., 2013)	valence-arousal (continuous)	345, 000	46
AFEW (Dhall et al., 2017)	7-basic face expressions	113,355	1809
FER2013	8-basic face eight basic facial expressions	28,709	Images
Aff-Wild	valence-arousal (continuous)	1,224,100	298
RAF-DB	7-basic face seven basic facial expressions	12,271	Images
AffectNet	7-basic face expressions	280,000	Images

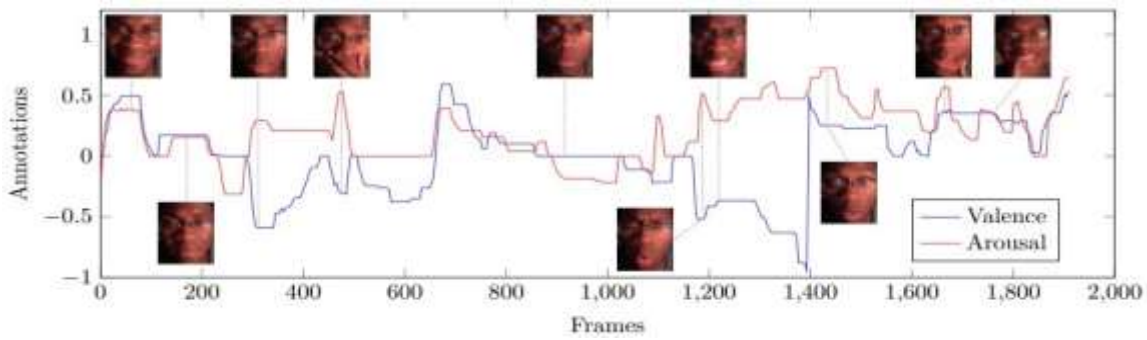

Figure 2.1 Throughout a video segment, annotations for valence and arousal and the corresponding frames demonstrate how Aff-Wild performs in the wild

2.3 Deep Learning Approaches

Recent research has demonstrated significant gains in several disciplines, such as speech recognition and photo identification, through deep learning (Afriyie et al., 2020; Jorge Martinez et al., 2021). With the help of deep learning algorithms, one of our objectives is to interpret timed facial expressions of emotions (Ahirwar et al., 2021). The proposed system can identify individuals through a camera. Analyze their expressions to simulate human emotions. Unlike methods that rely on created features, this system primarily utilizes facial expressions as its primary data source (Bhatt et al., 2020). It can distinguish between people. Recognize them based on their facial appearances.

Throughout this piece, we will look into the many components of the Deep CNN network that are used to classify the seven emotions. As an additional method for predicting valence and arousal levels, we investigate the CNN+RNN strategy. While some studies rely on deep convolutional neural networks, others incorporate a memory cell that collects frames to predict valence levels and current arousal. We will discuss both ways in detail and will emphasize the benefits and drawbacks of each.

The RELU activation function was employed alongside two convolution layers and two inception layers, each consisting of 1x1, 3x3, and 5x5 convolution layers, as Hussain and Al Balushi (2020) suggested to represent a network.

During the training procedure, a polynomial learning rate of baselr (1 iter/maxiter) that was comparable to 0.5 was assumed, as stated in their suggestion (Hussain and Al Balushi, 2020). An investigation revealed that their model had obtained an accuracy of 0.693, which was a significant finding.

On the other hand, Suryanarayana et al., (2021) produced a model for monitoring purposes based on the Hist eq approach. This model had an input transform layer, three pooling layers and a fully connected two-layer MLP. Suryanarayana et al., 2021 is the publication that brought this specific model to light. Compared to network models that claim an accuracy of 0.6667, the findings produced from this model were not only consistent but also remarkably consistent.

Zhang et al., (2019) state that a research team has presented a method for recognising facial expressions using the dataset. AFLW, Celeb Faces, and Kaggle were the three datasets used in the FER2013 model. These supplementary datasets included properties that were appropriately tagged. In order to establish a connection between the output and the dataset and use the aggregated features from all of these datasets, they developed a bridge layer. Regarding the identification of facial expressions, their system attained an accuracy measure of 0.71.

By analyzing the position and shape of facial landmarks, Devries and his colleagues (Devries et al., 2014) developed a method that enhances one's capacity to perceive facial expressions of emotion. Devries and colleagues' (2014) publication presented the methodology in question. The article "A strategy for analyzing the position and form of facial landmarks," which may be found online, describes this method. Their models include an output that uses the L2SVM activation function, three convolutional layers that are entirely linked, a ReLU hidden layer that is also fully connected, and three convolutional layers that are connected overall. In addition to that, their models consist of a total of three convolutional layers that are all coupled to one another. They employed various data augmentation techniques, including rotating, mirroring, zooming, and arbitrarily

reordering photos. Every one of these methods was applied. In addition, they reorganized the photographs in a completely arbitrary manner. Their technique resulted in an accuracy of 0.6721, which was reached by its use.

Table 2.2 Related work on the FER2013 data set

Models	Accuracy
Feng and Ren (Suryanarayena et al., 2021)	0.6667
Mollahosseini (Al Balushi and Hussain and Al Balushi, 2020)	0.693
Devries (Devries et al., 2014)	0.6721
Zhang (Zhang et al., 2019)	0.71

3. PROPOSED APPROACH

3.1 Introduction

In this part, we go into great depth on the training module and then proceed with training each model we used. The Adam optimizer, which we use for training and which we (Kingma and Ba 2014) integrate, incorporates two different gradient descent methodologies: Momentum consists of A strategy that is used to speed up the gradient descent process is one that makes use of the "exponentially weighted average" of the gradients. When averages are used, the method approaches the minima substantially more expediently.

In addition to that, we make use of a learning rate scheduler. By lowering the learning rate following the established plan, learning rate schedules aim to modify the learning rate during training.

To serve as a reference for the later construction of the Deep Learning models ResNet34 (He et al., 2016), MobileNet-V3 (Howard et al., 2017), and Wider ResNet50-2, we constructed a baseline model from the ground up. This model is intended to serve as our starting point. According to Zagoruyko and Komodakis (2016), this was used as a reference for the further development of these models. Following their initial training on the ImageNet data set, these three Deep CNN architectures were subjected to transfer learning to undergo further training later.

Transfer learning refers to applying knowledge and skills acquired in the past to new learning settings or situations that need problem-solving abilities. This technique is also known as "transfer learning." There is a possibility that this will take the shape of new learning settings or new situations. Consequently, it may be of the highest importance to explore the possibility of discovering parallels and similarities between the many learning processes that have occurred in the past and the content that is now being studied. To make it feasible for these models to be taught on our seven most

fundamental feelings, we tweak them so they can be trained. We can perform this for both the RAF-DB data set and the FER2013 data set. We are going to present a demonstration of our unique layers and talk about the influence that those layers have on the different models throughout every one of the training sessions.

This section displays each model's accuracy and cross-entropy loss we employ during the training and validation phases. The test set will be used to evaluate each of these models in the following section. We used ResNet34 to get the results that He et al., MobileNet-V3 (Howard et al., 2016) 2017), and Wider ResNet50-2 (Zagoruyko and Komodakis, 2016) on the FER2013 data set, we decided to only look at these Deep Learning architecture approaches on the other data set, RAF-DB, to see how well they handle the Emotion Recognition problem.

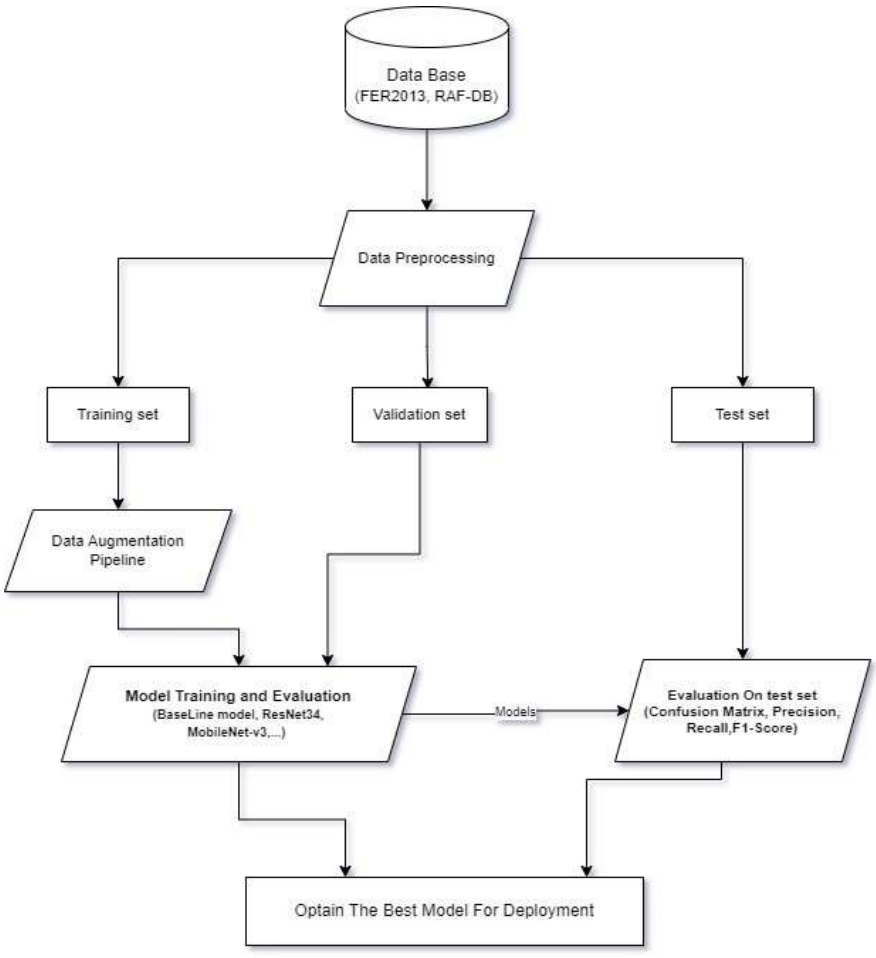

Figure 3.1 Optimal Model Pipeline

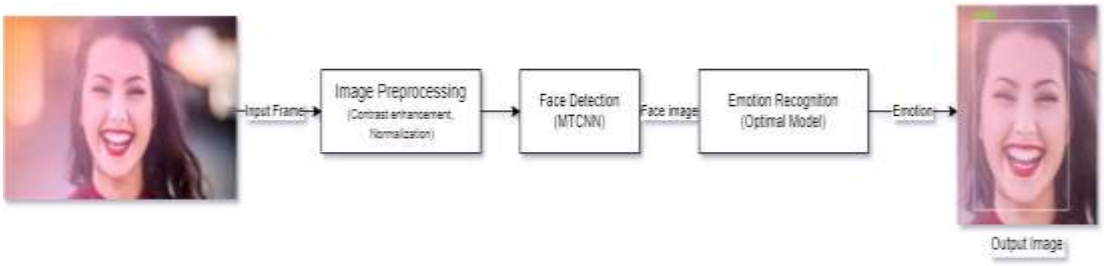

Figure 3.2 The deployment pipeline

3.2 Methods and Structures For Deep Learning and Convolutional Neural Networks (CNNs)

3.2.1 Convolutional neural network

(CNN) is a widely utilized deep learning technology in several applications requiring picture recognition and classification. This kind of deep neural network requires the smallest amount of pre-processing of any other kind since the algorithm learns all filters, and there is no hand-crafting of features. Because sending the image into the network one pixel at a time makes it more challenging to recognize ambiguous patterns in the picture, it is transmitted into the network in discrete chunks rather than one pixel at a time. A CNN has three unique levels: input, output, and multiple hidden layers. A layer in the middle separates these levels. Layers with completely connected connections, layers with normalized values, layers with maximum and average pooling, and layers with normalized values are hidden in CNN layers. Other examples of hidden layers are the Convolutional and Pooling layers (Pashine et al., 2021). A CNN employs a kernel, often called a filter, a compact set of weights to analyze the input image and extract diverse information. The convolution operation entails moving the kernel over the input image and calculating the scalar product of the kernel values and the corresponding pixel values in the input. Observing CNN shows that the vertical and horizontal dimensions are diminishing, even though the number of channels is expanding. Ultimately, the created column matrix predicts the result (Udofia 2018).

CONVOLUTIONAL neural networks (CNNs), also called convents, have demonstrated their efficacy in tasks (Minaee et al., 2019; Wayman et al., 2005). Each layer of a CNN

contains a set of filters that combine channel-wise information within receptive fields. These filters provide insights into the connections among neighbouring patterns across input channels. By incorporating linear activation functions and down-sampling operators, CNNs can generate image representations capable of capturing hierarchical patterns and achieving global theoretical receptive fields.

This allows CNN to achieve global theoretical receptive fields. This objective has been fulfilled. One of the key aims of research that is being carried out in computer vision is the quest for more effective representations that can isolate the components of an image that are most important to a particular job to allow improved performance levels. Now that we have reached this point, one of the most essential things we need to do in the next stage is build new architectures for neural networks. These structures should be able to be utilized as a family of models to tackle a wide range of vision problems. Recent research has shown that convolutional neural networks (CNNs) may provide more accurate representations of data if they are equipped with learning mechanisms on the node level. These learning processes are critical since they assist in capturing spatial correlations between data sets. The Inception family of designs is responsible for bringing widespread attention to one method that integrates processes operating on several scales into network modules (Hjelms and Low 2001, Zhang and Zhang 2010). This tactic is used to enhance performance. There has been an increase in research done to understand geographical links better and include spatial awareness in the network's structure. This is being done both in the United States and internationally.

Recent studies by Simonyan and Zisserman (2014) on VGGNets and Szegedy et al., (2017) on Inception models have shown that the quality of the representations that a network learns can be significantly improved by increasing its depth. This improvement is further supported by the introduction of Batch Normalization (BN) (Ioffe and Szegedy 2015), which helps stabilize the learning process in networks and creates optimization surfaces by managing input distributions at each layer. Another breakthrough came with ResNets, where identity-based skip connections allowed for training more powerful networks (He et al., 2016). Additionally, highway networks [47] incorporated a gating mechanism to control information flow through connections.

Building on these advancements, researchers have reformulated connections between network layers, leading to promising enhancements in both learning and representation capabilities.

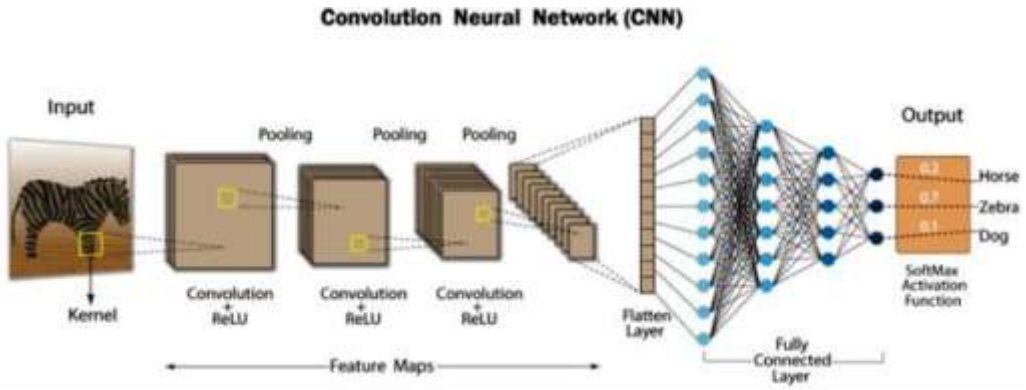

Figure 3.3 Information about the architecture of an image-recognition convolutional neural network (CNN) (O'Shea and Nash 2015)

3.2.2 Mobilenet-V3

Over time, there has been a significant improvement in the efficiency of the building blocks that are used for models. The depthwise convolutions developed by MobileNetV1 (Howard et al., 2017) were an efficient alternative to the conventional convolution layers. The filtering process is separated from the feature creation process using this method, ultimately resulting in the factorization of traditional convolutions. Both depthwise convolutions, which are used for filtering, and heavier 1x1 pointwise convolutions, which are used for feature synthesis, are included in the two layers that make up depthwise separable convolutions.

MobileNetV2 significantly improved layer structures by including the bottleneck and inverted residual structures (Howard et al., 2018). This improvement was accomplished by using the rank nature of the issue. A 1x1 expansion convolution, a depthwise convolution, and a 1x1 projection layer are all components of this construction, as seen in Figure 3.4. In order for there to be a residual connection, Input and output must have the same number of channels. This structure grows internally to a dimensional feature

space while preserving a compact representation at the input and output to improve the per-channel transformations.

By incorporating lightweight attention modules into the bottleneck structure, MnasNet (Tan et al., 2019) made substantial breakthroughs in the architecture of MobileNetV2. The groundwork for these attention modules was laid with the concepts of squeeze and excitement. It is essential to note the presence of the squeeze and excitation module, which is somewhat distinct from the ResNet-based modules that were suggested in the sentence before this one (Hu et al., 2018). Figure 3.5 demonstrates that the module is positioned inside the expansion to follow the depthwise filters. This can be observed by looking at the diagram. This action guarantees that the representation accountable for the most significant amount of data receives the attention it deserves.

To develop the most efficient models that are possible and accessible for MobileNetV3, the writers of MobileNetV3 make use of these layers while they are creating parts. The layers have been updated to reflect that the nonlinearities in Swish that required correcting have been resolved. This update was done in order to reflect the patch that was applied. In place of the sigmoid, which is used by squeezing, excitation, and the swish nonlinearity, they use the hard sigmoid. On the other hand, the sigmoid may be computationally expensive and difficult to maintain precision in fixed point arithmetic. Squeezing, excitation, and swish nonlinearity are used to exploit these nonlinear functions.

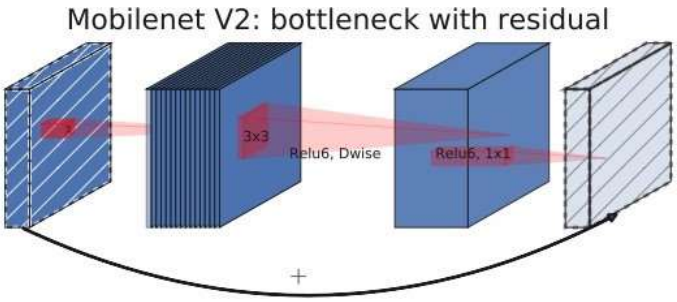

Figure 3.4 (Inverted Residual and Linear Bottleneck) MobileNetV2 (Howard et al. 2018) layer The narrow input and output (bottleneck) of each block is followed by expansion into a much higher-dimensional space and projection to the output, which is both linear. Bo

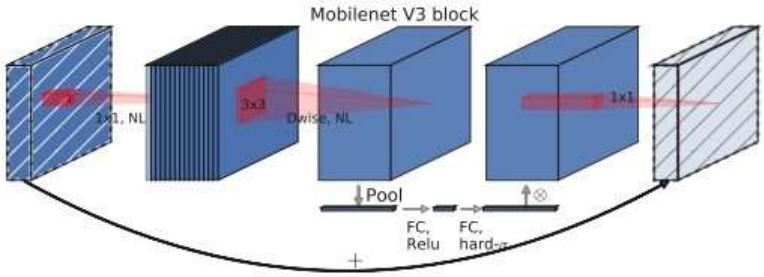

Figure 3.5 Squeeze-and-Excite plus MobileNetV2 (Hu et al., 2018) They apply the squeeze and excite in the residual layer, in contrast to (Hu et al., 2018). Depending on the layer, they employ various nonlinearities

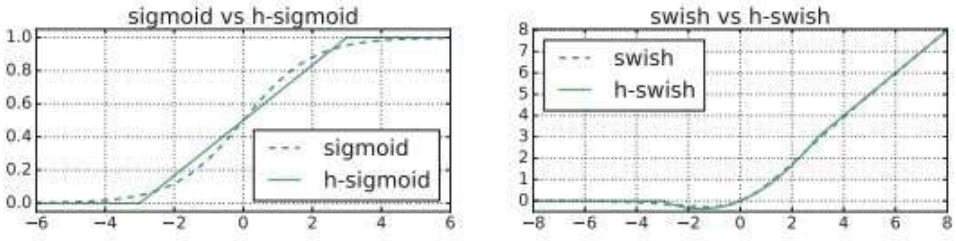

Figure 3.6 Sigmoid and swish nonlinearities and their “hard” counterparts

Table 3.1 MobileNetV3-Large specification SE indicates whether a Squeeze-And-Excite exists in that block. The use of nonlinearity is indicated by the letter NL. In this case, HS stands for h-swish and RE for ReLU. NBN indicates no batch normalization.

Input	Layer	Expansion Size	Output size	SE	NL	s
$112^2 \times 16$	bneck,3x3	16	16		RE	1
$112^2 \times 16$	bneck,3x3	64	24		RE	2
$56^2 \times 24$	bneck,3x3	72	24		RE	1
$56^2 \times 24$	bneck,5x5	72	40		RE	2
$28^2 \times 40$	bneck,5x5	120	40		RE	1
$28^2 \times 40$	bneck,5x5	120	40		RE	1
$28^2 \times 40$	bneck,3x3	240	80		HS	2
$14^2 \times 80$	bneck,3x3	200	80		HS	1
$14^2 \times 80$	bneck,3x3	184	80		HS	1
$14^2 \times 80$	bneck,3x3	184	80		HS	1
$14^2 \times 80$	bneck,3x3	480	112		HS	1
$14^2 \times 112$	bneck,3x3	672	112		HS	1
$14^2 \times 112$	bneck,5x5	672	160		HS	2
$7^2 \times 160$	bneck,5x5	960	160		HS	1
$7^2 \times 160$	bneck,5x5	960	160		HS	1

Table 3.1 MobileNetV3-Large specification SE indicates whether a Squeeze-And-Excite exists in that block. The use of nonlinearity is indicated by the letter NL. In this case, HS stands for h-swish and RE for ReLU. NBN indicates no batch normalization (continue)

Input	Layer	Expansion Size	Output size	SE	NL	s
$7^2 \times 160$	conv2d,1x1	-	960		HS	1
$7^2 \times 960$	pool,7x7	-	-		-	1
$1^2 \times 960$	conv2d,1x1,NBN	-	1280		HS	1
$1^2 \times 1280$	conv2d,1x1,NBN	-	K		-	1

3.2.3 Resnet

ResNet was developed by Kaiming He and his colleagues in 2016 (He et al., 2016). Using a residual learning approach, it is possible to train deeper networks, which are notoriously difficult to train in practice. It was decided to rebuild the network layers so that they could learn residual functions, taking into consideration the layer inputs.

Based on the inquiry's findings, it was determined that deeper networks founded on residual learning could achieve higher accuracy and better optimization. There is a need for further citations. There is a need for further citations. According to the research results (Simonyan and Zisserman 2014, Szegedy et al., 2015), it is required to significantly broaden the network's scope at deeper levels to achieve greater performance levels. Deeper networks are notoriously difficult to train, but adding extra layers could increase learning.

Additionally, at some point during deep network convergence, accuracy will reach a point where it will get saturated, and then it will begin to fall rapidly from that point on. This change will occur at some point in time. The problem of accuracy erosion that occurs in deeper networks may be solved by using residual learning in these networks, as it provides a solution to the problem. This solution is because residual learning provides a solution to the problem.

Simple networks can rapidly learn the most desirable mapping by stacking numerous layers on top of one another. On the other hand, in residual networks, the layers are layered to learn a residual mapping structure. The mapping function, denoted by the notation H(x), is compatible with several models.

Strata-like layers are organized in layers; according to the theory that drives residual learning, if several nonlinear layers can estimate a hard mapping function asymptotically, those layers should also be able to do the same for the residual function, represented by the symbol F. This approach is because the residual function is the function that is being estimated. The term "difficult mapping function" (x) refers to this particular aspect of the assignment. By doing the following, it is possible to accomplish the fundamental mapping:

$$H(x) = F(x) + x \tag{1}$$

Instead of learning the basic function H(x), which they could have been able to achieve, the stacked layers educate themselves on the residual function F instead (x). This change in functions is more important than learning the fundamental function. According to this method, the optimization of the residual mapping function should require a great deal less complexity than the optimization of the initial function did. This simplification is because the residual mapping function is smaller than the original. In contrast to the approach of estimating the identity mapping by stacking a few non-linear layers, the method of estimating the identity mapping is better because it allows the residual to be readily reduced to zero. For example, if there is an identity mapping, the residual can quickly be reduced to zero. A discovery was made while computing the residual of the function, which revealed that the original mapping function was H(x) = F(x) + x. Following the estimation of the function's residual, this was found out. Following the completion of an estimate of the function, this was discovered. The mapping function F(x) + x, encoded into a feed-forward neural network as a residual shortcut link, is used to carry out this element-wise addition. This addition is accomplished by using the mapping function.

Consequently, element-wise addition may be done by employing this. The data that is used by the network originates from a particular source. It is possible to draw parallels between these links in a residual network and identity mapping in the natural world. After that, the output reapplies to the layers already stacked in the phase. These connections have no impact on the intricate architecture of the networks or the characteristics they display in any way, shape, or form. Furthermore, in order to expedite the training process for these residual networks, it is possible to make use of back-propagation that is based on SGD.

Figure 3.8 depicts the basic concept developed for the ResNet 18 system. The network comprises 18 layers, including 17 layers, one linked layer, and an extra softmax layer for classification. If the output feature map is identical to the dimensions of the 3x3 filters used in the layers, the design guarantees that those layers will include several filters. On the other hand, if the feature map produced is half as large as before, then each layer will have filters appearing as often. Convolutional layers with two striations are used to minimize the data amount.

For the very lowest two levels of the structure, the softmax layer, the fully connected, and the average pooling layer are the layers that make up the structure. Lastly, a layer is ultimately connected to the structure. The remaining shortcut links between floors have been implemented throughout the network to finish the layout. This link was done across the entire network. Regarding forming relationships with other individuals, two distinct approaches may be used. When there is no discernible difference in size between the input and the output, the first kind of connection, represented by solid lines, is used to establish a connection between the associated components. Many types of connections, depicted by dashed lines, are used to accommodate the growth in size. Despite having a stride of 2, this connection type continues to engage in identity mapping; however, it does so by padding more significant connections.

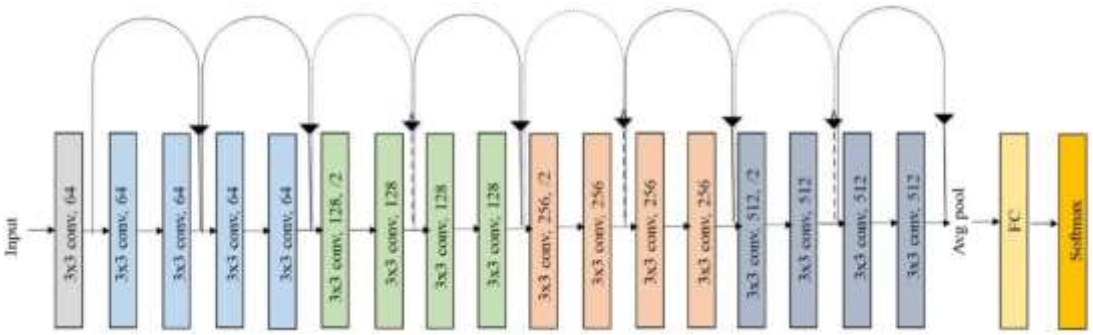

Figure 3.7 Overview of ResNet-18 architecture (Ramzan et al., 2020)

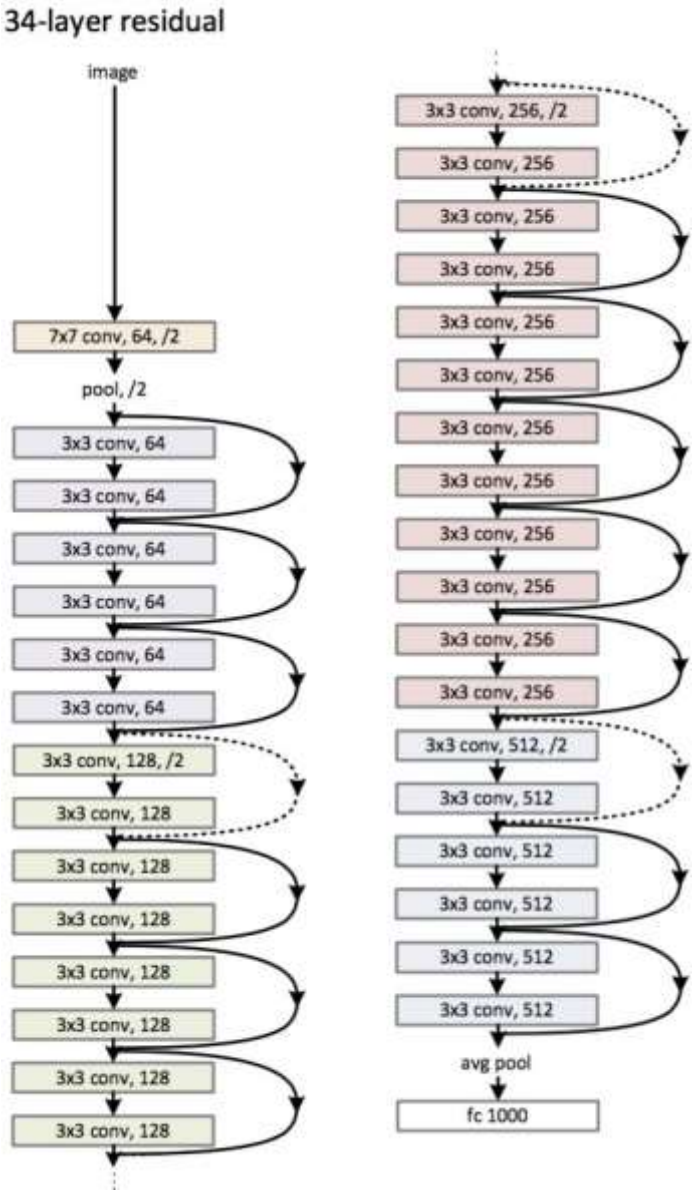

Figure3.8 Overview of ResNet-34 architecture (He et al., 2016)

3.2.4 Wide residual networks (WRNs)

WRNS (Wide Residual Networks) are being demonstrated this time. The network may be made shallower while retaining or even increasing its degree of accuracy by increasing the Residual Network (ResNet).

- Layer count can be decreased.
- An even less amount of time may be devoted to training.

Less in-depth models may require exponentially more components. The people who developed residual networks sought to make them as thin as was reasonably conceivable to have a larger depth and fewer parameters. This simplification was done in order to make them more efficient. They also included a "bottleneck" block, which makes ResNet blocks even more constrained in their width.

Although it is technically possible for the gradient to flow through the weights of the blocks in a network and allow it to avoid learning anything during training, there is still a chance that Only a few blocks learn meaningful representations or a large number of those blocks contribute very little information to the overall goal. Both of these situations can occur even though nothing prevents the gradient from flowing through the block weights of the network. Despite this lack of obstruction for flow through the block weights, each scenario remains a possibility. This problem has been referred to as "diminishing feature reuse" by Zagoruyko and Komodakis (2016).

In WRNs, a wide range of parameters, such as the design of the ResNet block and its depth (deepening factor l) and width, are analyzed. For example, in WRNs, the design of the ResNet block is examined (widening factor k).

When k equals 1, it has the same width as ResNet. Although k is a positive number, it is much more extensive than ResNet by a factor of k.

The symbol d denotes the depth of the WRN, and the widening factor is denoted by the letter k in the notation WRN-d-k. In our investigation, we use a model known as the WRN-50-2, which specifies that the network has fifty layers and that its width is twice as large as that of the ResNet50 model.

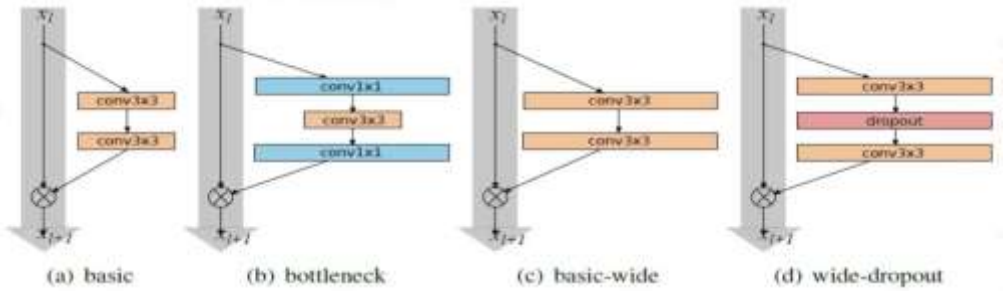

Figure 3.9 Various ResNet Blocks (Zagoruyko and Komodakis 2016)

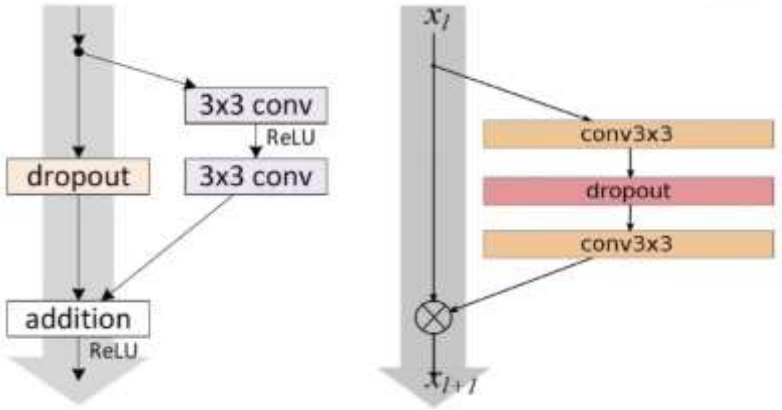

Figure 3.10 (Left) Dropout in Original ResNet (Right) Dropout in WRNs (Zagoruyko and Komodakis 2016)

3.3 Face Detection

Face detection and matching play a role in our pipeline. In this step, we will discuss two approaches to extracting all faces from images. The first is the cascade approach, where several well-orchestrated networks are used to improve the detection accuracy (Zhang et al., 2016). The second approach is called a single shot detector, e.g. YOLO, which uses

a snapshot to detect objects in an image using a single snapshot and divides the image into a grid of boxes.

Face detection is one of the component pieces that may be extracted from object-class detection. Object-class detection can be disassembled into its parts. Object-class detection is the process of detecting the locations and sizes of all entities in an image that belong to a certain class. This detection may be accomplished by analyzing the image in question. In this particular illustration, the example class is a human face. Finding the positions of these entities and their dimensions is the objective of the object-class identification process. During this first phase of our processing pipeline, our primary emphasis is locating and excluding as many faces from each image as possible. In the first processing stage, we check each video frame using face detection (Minaee et al., 2021).

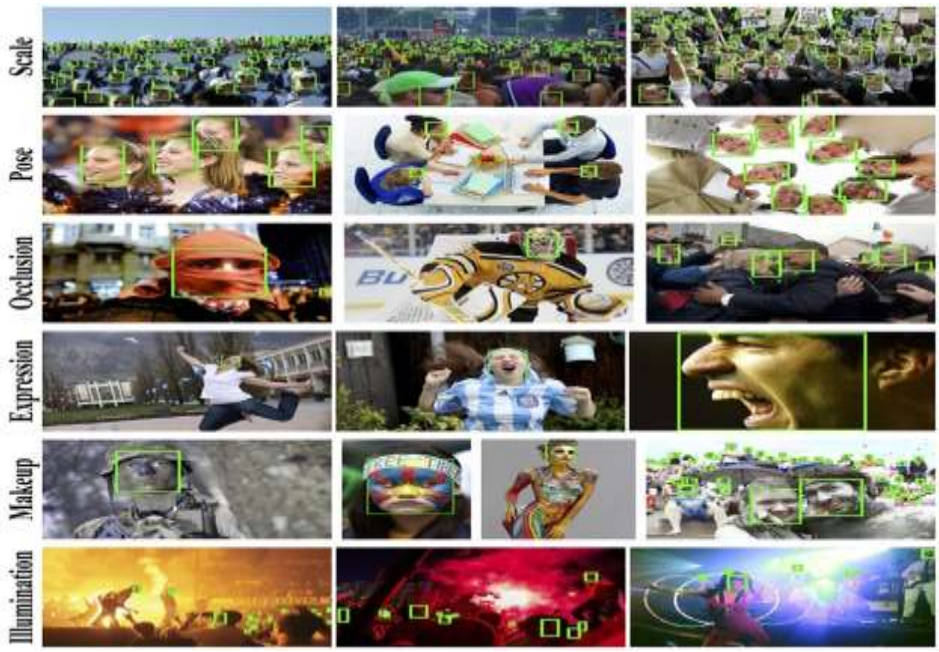

Figure 3.11 A set of images from “Wider-Face” face detection data sets Reviews the different factors of difference (Yang et al., 2016)

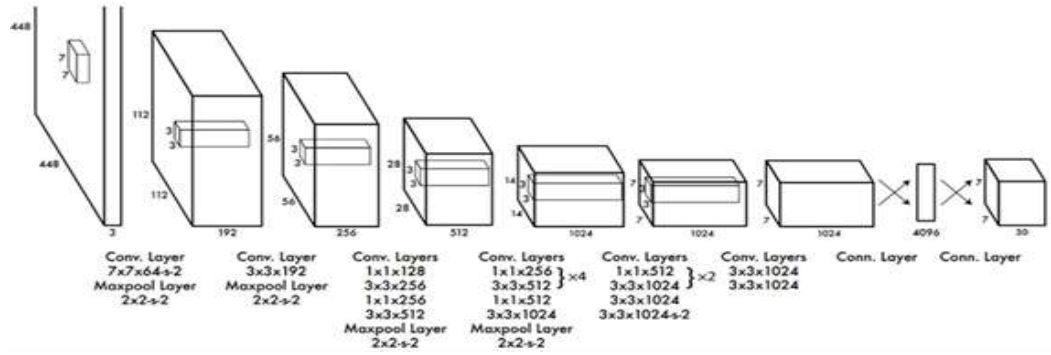

Figure 3.12 Random images from the validation set with its prediction and real labels (Zhang et al., 2016)

MTCNN, known as Multi-task Cascaded Convolutional Networks, is an approach we plan to incorporate into our pipeline due to its accuracy on our test set. The MTCNN consists of three stages:

• Stage 1: Proposal Network, or P Net, a CNN model. This step aims to perform bounding box regression and find candidate windows in an image (Minaee et al., 2021).

• Stage 2: R Net, another CNN model called Refine Network, is employed to filter out positives further.

• Stage 3, similar to the second stage, aims to provide a more detailed description of the face. The network outputs the positions of four landmarks: x coordinate, y coordinate, height and width.

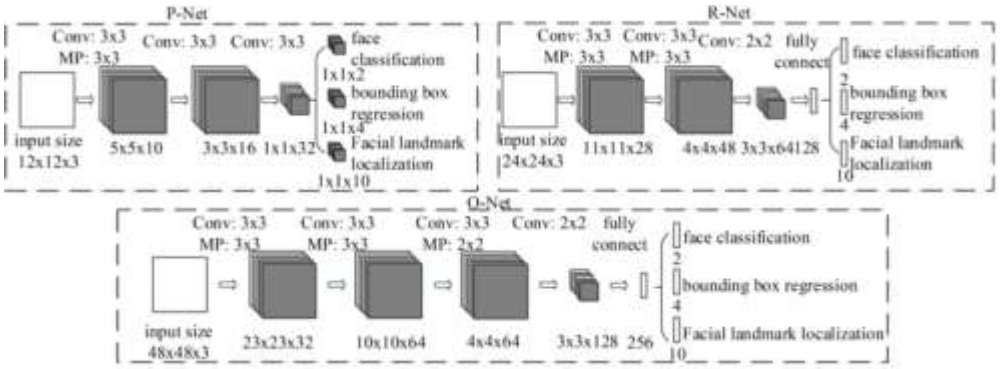

Figure 3.13 In the P-Net, R-Net, and O-Net designs, "MP" stands for maximum pooling and "Conv" for convolution. The step size is two in pooling and convolution, respectively [(Zhang et al.,)]. 2016)

3.4 Data Augmentation

We need to expand our dataset to tackle the issue of overfitting. The current amount of data we have can be increased by making modifications. The goal is to achieve results that resemble the variations observed in images or videos when capturing them.

Data augmentation techniques involve altering the training data to impact its array representation while keeping the label intact. These approaches are employed to enhance the accuracy of machine learning models. Various augmentations are commonly utilised, such as flipping vertically, randomly cropping, introducing colour variations, translating, rotating, and more.

By making just a few modifications to our training data, we can quickly increase the number of training instances by two or four times, allowing us to construct a very resilient model. The results of these augmentation methods are shown in Figures 3.14 and 3.15, which compare and contrast examples from the two different data sets. An example of data enhancement using maximum ranges may be seen here.

- Random Rotation ranges from [-10,10] degrees.
- Adjust brightness
- Crop and padding
- Width Shift
- Height Shift
- Horizontal Flip with 50% probability.

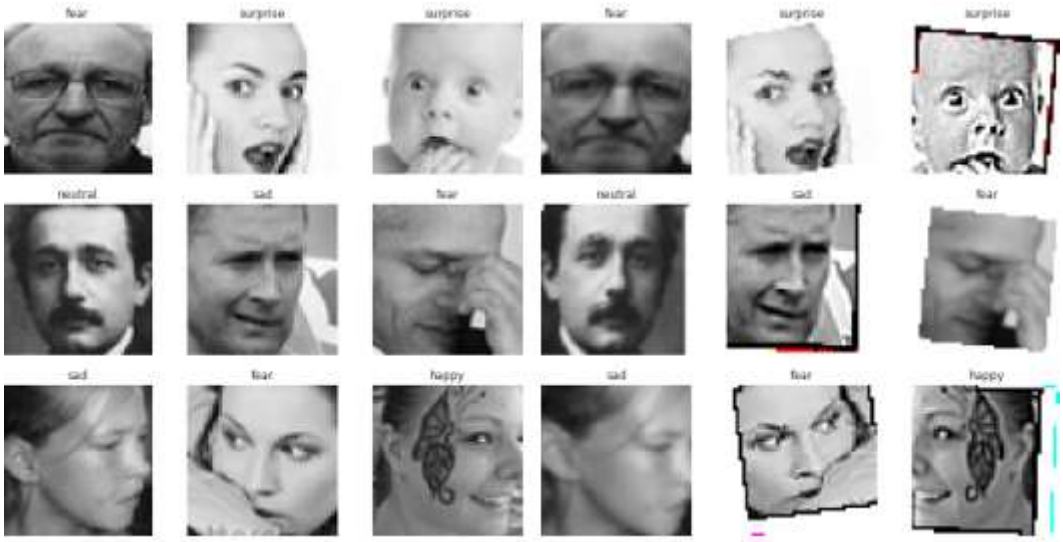

Figure 3.14 FER2013 data set before and after we used our pipeline for augmentation. The samples on the left were taken before using our augmentation process, while the ones on the right were taken following the application of the pipeline

Figure 3.15 RAF-DB data set before and after we used our pipeline for augmentation. The samples on the left were taken before using our augmentation process, while the ones on the right were taken following the application of the pipeline

3.5 Evaluation Metrics

When dealing with these evaluation measures, the main objective is to determine how successfully a machine-learning model will perform when provided with new data.

When evaluating classification models for data sets that are thought to be balanced, metrics such as accuracy, precision, and recall are appropriate approaches to use as evaluation tools. On the other hand, if the data are not uniformly distributed, the model's performance may be better evaluated using other approaches such as ROC/AUC.

The receiver operating characteristic (ROC) curve is more than just a single number since it is a complete curve that gives considerable information about the behaviour of the classifier. This definition means that a single number cannot represent the ROC curve. In addition, making a speedy comparison of many distinct ROC curves to one another requires some time.

3.5.1 Accuracy

Accuracy is determined by how the classifier makes predictions. It can be characterized as the ratio of predictions to total predictions.

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \quad (2)$$

" There is a possibility that the impression that a model is doing well is created when the model claims to have an accuracy rate of 99%. On the other hand, there are situations in which this is not the case, and it should be considered misleading. The following is a concrete example to illustrate this idea immediately. Accuracy is beneficial when the targeting class is well-balanced, but it is not an appropriate choice for courses that are out of whack with one another. Imagine if our training data consisted of 99 per cent of pictures of the dog but just one per cent of the cat. If this were the case, our model would always correctly predict the dog, giving us an accuracy of 99%. Data is never balanced, as shown by the prevalence of fraudulent activities such as credit card fraud, spam emails, and inaccurate medical diagnoses. Suppose we want a fuller view of the model assessment and conduct a better model evaluation. In that case, we need to consider the metrics of recall and accuracy and any additional metrics that may be applicable.

Our data sets do not suffer from unbalanced data, although there is a tiny variation between the goal labels for positive and negative outcomes. Therefore, to better understand how our models function, we consider accuracy, recall, F1-score, and AUC metrics. Additionally, if we want to tweak the precision-recall trade-off, it may be best if we change it. When accuracy is boosted, it will have a negative impact on recall and vice versa. This kind of compromise is referred to as the precision-recall trade-off. Since the classifier's performance varies depending on the threshold value, it is possible to alter both the positive and negative predictions by adjusting the threshold value.

3.5.2 Precision

Precision indicates the proportion of predicted situations that turned out to be positive. When false positives carry consequences, negative precision becomes beneficial. E-commerce websites, music or video recommendation systems and other applications heavily rely on precision to avoid outcomes that may drive customers away, which is detrimental to business success. To determine the precision of a label, Calculate the ratio of the total number of genuine positive outcomes to the total number of anticipated positive results.

$$P = \frac{TP}{TP + FP} \quad (3)$$

3.5.3 Recall

Prediction accuracy in determining outcomes is measured by precision. When false positives are more concerning than negatives, they become beneficial. Precision is vital in applications like music or video recommendation systems, e-commerce websites, and others where inaccurate results can turn customers away and hurt businesses. On the other hand, recall Determines the positive ratio of all actual positives for a given label.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

3.5.4 F1 score

It is a measure commonly used to evaluate the performance of a classification model, especially when there is an imbalance between classes. It is considered a harmonious means of knowing accuracy and recall.

$$F1 = 2 \times \frac{\textbf{Precision} \times \textbf{Recall}}{\textbf{Precision} + \textbf{Recall}} \tag{5}$$

The F1 score penalizes more high values. In the following circumstances, the F1 Score may be useful:

- When FP and FN are equally costly.
- Adding more data does not effectively change the outcome.
- True Negative is high.

3.5.5 Confusion matrix

- Figure 3.16 depicts a binary classification problem typically involves Positive and negative classes. Now, let us delve into the metrics of the Confusion Matrix.
- Positive (TP): The number of cases correctly predicted as positive.
- True Negative (TN): The number of cases correctly predicted as negative.
- False positive (FP): The number of cases incorrectly predicted as positive. Also known as type I error.
- False Negative (FN): The number of instances incorrectly predicted as negative, also known as a Type II error.

- The confusion matrix is usually used as an evaluation criterion for a machine learning model. It provides an effective way to evaluate the performance of models.

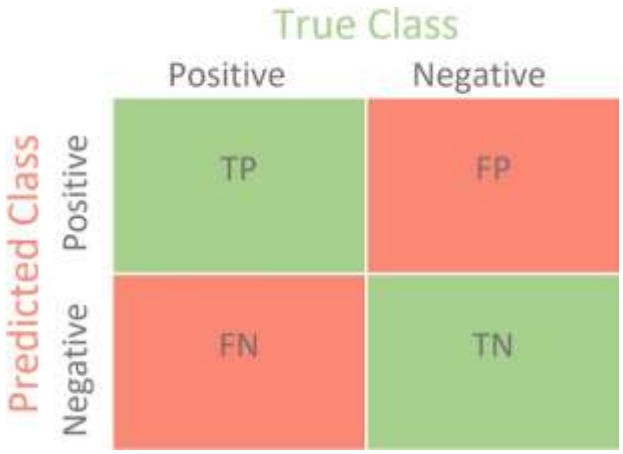

Figure 3.16 Confusing Matrix for Classifying Binary Data

4. TRAINING AND EXPERIMENT

4.1 Introduction

In this part, we thoroughly describe the training module and go through the process of training each model we used independently of the other. For training purposes, we use the Adam optimizer (Kingma and Ba 2014). The Adam optimizer is defined by its mix of two distinct gradient descent techniques. This method accelerates the gradient descent by incorporating an exponentially weighted gradient average. Momentum: The algorithm exhibits a significant swifter movement towards the lowest threshold value when employing averages.

In addition, we employ a learning rate scheduler approach. Learning rate schedules aim to modify the learning rate throughout training by decreasing it based on a predetermined timetable.

In order to serve as a guide for the later construction of the Deep Learning models ResNet34 (He et al., 2016), MobileNet-V3 (Howard et al., 2017), and Wider ResNet50-2, a baseline model was built from the ground up. This baseline model was done to serve as a reference for the subsequent development of these models (Zagoruyko and Komodakis, 2016). Transfer learning was used to train these three Deep CNN architectures further after being pre-trained on the ImageNet data set.

Transfer learning refers to applying acquired knowledge and skills in learning or problem-solving scenarios. The similarities and analogies between present learning content and processes can be significant.

We modify these models so that they can be trained on our seven essential emotions. We accomplish this for both the FER2013 data set and the RAF-DB data set. Throughout each of the training sessions, we will provide a demonstration of our distinctive layers and discuss the impact that those layers have on the various models.

This section shows the training and validation accuracy and cross-entropy loss of each model we use, and in the next section, we evaluate all these models on the test set. The result we obtained using ResNet34 (He et al., 2016), MobileNet-V3(Howard et al., 2017), and Wider ResNet50-2 (Zagoruyko and Komodakis, 2016) on the FER2013 data set to make us only consider these approaches on the other data set RAF-DB.

4.2 Dataset Analysis

In this part, we will concentrate only on the face and attempt to go backwards to create more complicated models. The human face is the portion of our body that holds the most information and is the most significant in recognizing facial gestures. We will use a data set from Kaggle that consists of grayscale photos of size 48*48 to get started. This picture contains much noise, and the data set is challenging to work with since it is difficult to achieve high precision.

The FER2013 data collection comprises grey-scale images of faces that are 48 pixels wide and 48 pixels tall. The facial features have been automatically catalogued so that they are all nearly in the same place and take up about the same amount of space as one another.

Each face must be assigned to one of seven categories, where 0 means anger, 1 means disgust, 2 means fear, 3 means happiness, four means sadness, five means surprise, and 6 means neutral. The training set contains 28,709 instances, and the public test set contains 3,589.

The Real-world Affective Faces Database (RAF-DB) contains over 30,000 face pictures with basic or complicated expression annotations. These facial photos may be accessed in the database (RAFDB). Only the 12,271 responses annotated with the core emotions were utilized for the training component of the study.

The distribution of the seven labels in the data set is similar to what is shown in figures 4.1 and 4.2 below. Compared to the other labels in the FER2013 data set, the Disgust label has a much smaller number of associated pictures, which may cause some difficulties during the learning process. In RAF-DB, fear labels have fewer photos than other labels, yet both data sets include a wealth of images for the joyful label, as seen in the two figures.

Figures 4.3 and 4.4 provide sample data from the two data sets, FER2013 with grayscale photos and RAF-DB with RGB images. These samples were offered to understand better how our data appears.

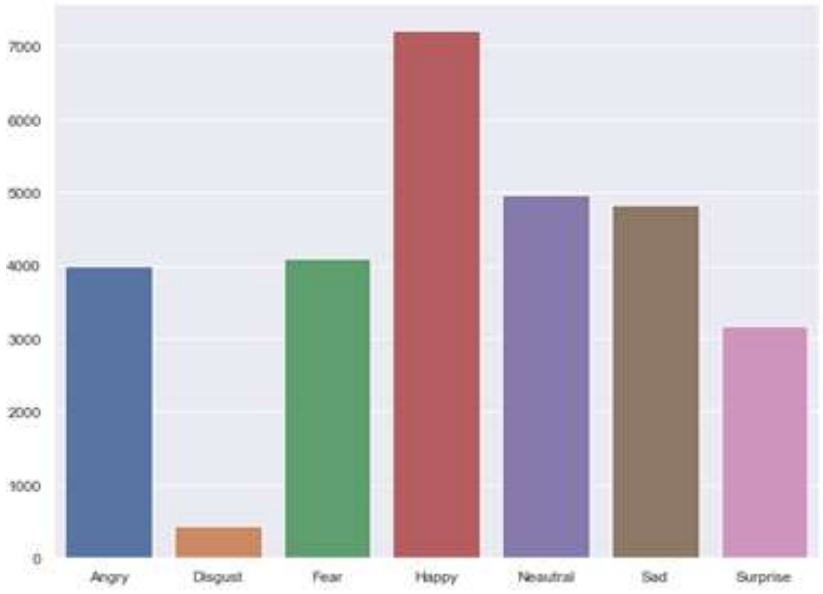

Figure 4.1 A bar chart showing the fer 2013 dataset on Kaggle (Label vs. Total training sample size)

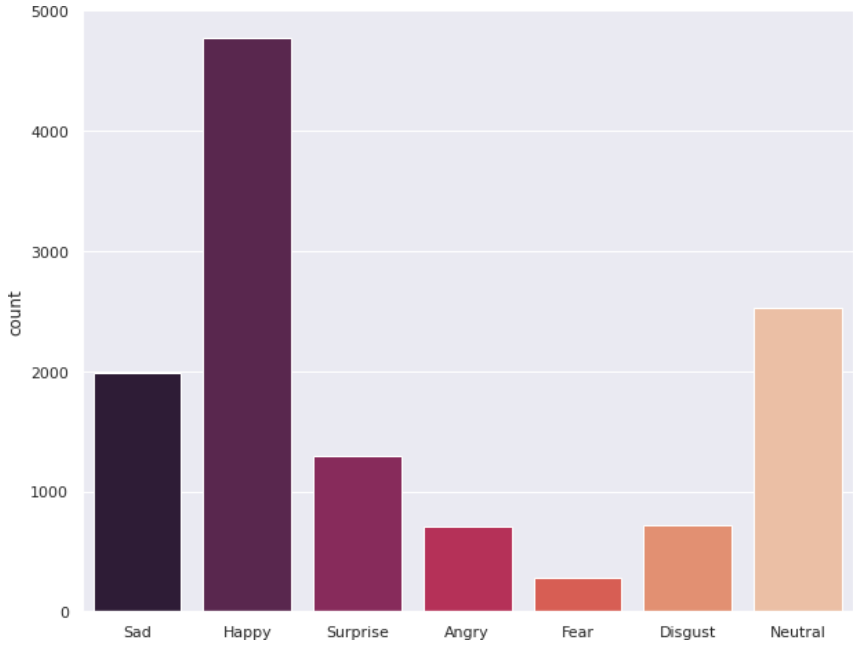

Figure 4.2 A bar chart showing the RAF-db dataset on Kaggle (Label vs Total training sample size)

Figure 4.3 Some examples of grayscale images from the FER2013 dataset

Figure 4.4 Some RGB image samples from the RAF-DB data set

4.3 Build a Baseline Model

In this section, our primary focus is on building a starting point model from scratch. This model will be a reference for creating the Deep Learning models ResNet34 MobileNet V3 and Wider ResNet50 2.

Training neural networks can be challenging. To overcome this, we adopt a learning framework that helps train these networks effectively. As the network depth increases, the accuracy eventually reaches a saturation point. Then, it declines rapidly. Therefore, simply adding layers does not necessarily improve accuracy. However, the residual framework facilitates training layers (He et al., 2016).

This block serves as the foundation for our architecture. The Keras plot of the model's architecture describes its connections and layer count.

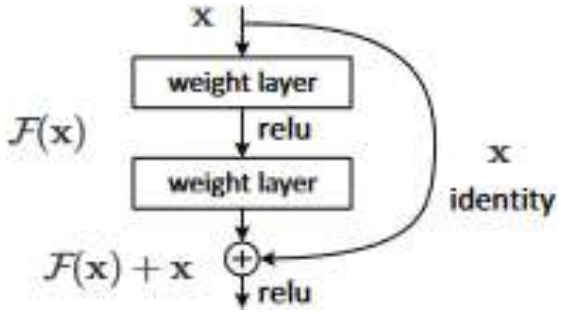

Figure 4.5 Residual learning: a building block (Kaiming al. 2015)

Layer (type)	Output Shape	Param #
conv2d_24 (Conv2D)	(None, 24, 24, 64)	1600
batch_normalization_24 (Batc	(None, 24, 24, 64)	256
activation_1 (Activation)	(None, 24, 24, 64)	0
max_pooling2d_1 (MaxPooling2	(None, 12, 12, 64)	0
residual_unit_10 (ResidualUn	(None, 6, 6, 32)	30080
residual_unit_11 (ResidualUn	(None, 6, 6, 32)	18688
residual_unit_12 (ResidualUn	(None, 3, 3, 64)	58112
residual_unit_13 (ResidualUn	(None, 3, 3, 64)	74240
residual_unit_14 (ResidualUn	(None, 3, 3, 64)	74240
residual_unit_15 (ResidualUn	(None, 2, 2, 128)	230912
residual_unit_16 (ResidualUn	(None, 2, 2, 128)	295936
residual_unit_17 (ResidualUn	(None, 2, 2, 128)	295936
residual_unit_18 (ResidualUn	(None, 2, 2, 128)	295936
residual_unit_19 (ResidualUn	(None, 2, 2, 128)	295936
global_average_pooling2d_1 (	(None, 128)	0
flatten_1 (Flatten)	(None, 128)	0
dense_1 (Dense)	(None, 7)	903

Total params: 1,672,775
Trainable params: 1,668,615
Non-trainable params: 4,160

Figure 4.6 Model Architecture

4.3.1 Training the baseline model

We incorporate hyperparameter tuning as part of our process, where we carefully explore parameters to optimize the models' performance. We trained our model for 300 epochs, with a learning rate we set at 0.001. Considering that our dataset consists of grayscale images sized 48x48 pixels, achieving an accuracy of 60.05 per cent on the validation set is entirely satisfactory. Throughout the 300 epochs, we observe improvements in accuracy, as demonstrated in Figure 4.7, while Figure 4.8 displays the changes in loss values for both the training and validation sets. Additionally, Figure 4.9 showcases our measurements using the AUC metric (area under the curve). To enhance training efficiency, we employ a learning rate scheduler, "Reduce on Plateau ", depicted in Figure 4.10. This scheduler monitors a metric. Decreases the learning rate if there is no progress after a predefined number of epochs, referred to as 'patience.' On a CPU, each epoch requires two minutes for training.

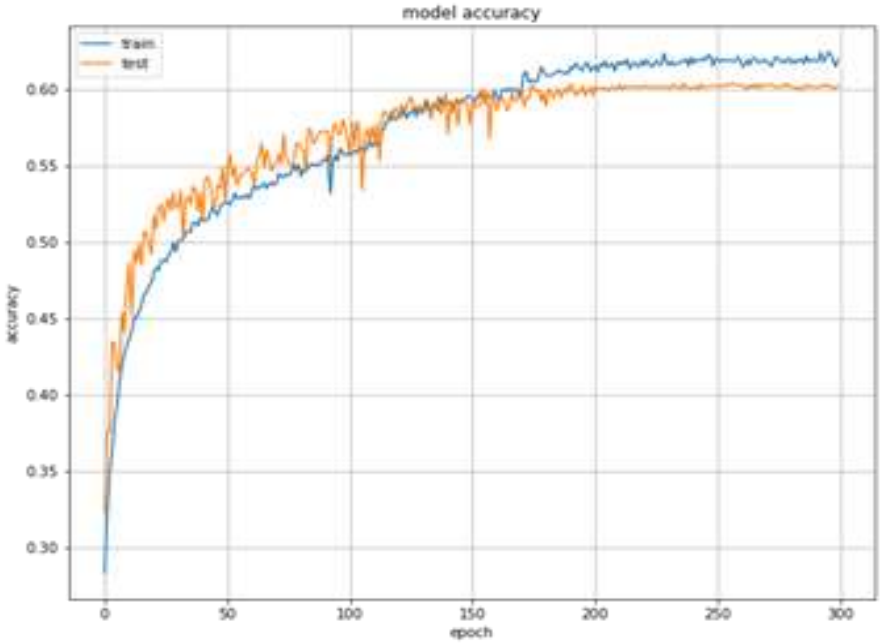

Figure 4.7 The training set's accuracy for the baseline model is represented by the colour blue, whereas the colour orange represents the accuracy of the validation set

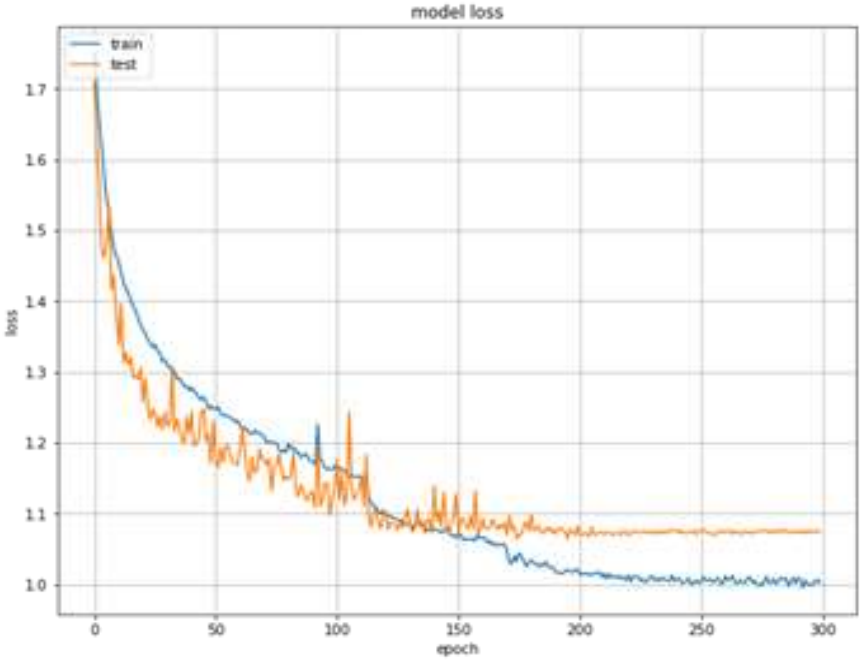

Figure 4.8 The baseline model's loss from the training and validation sets

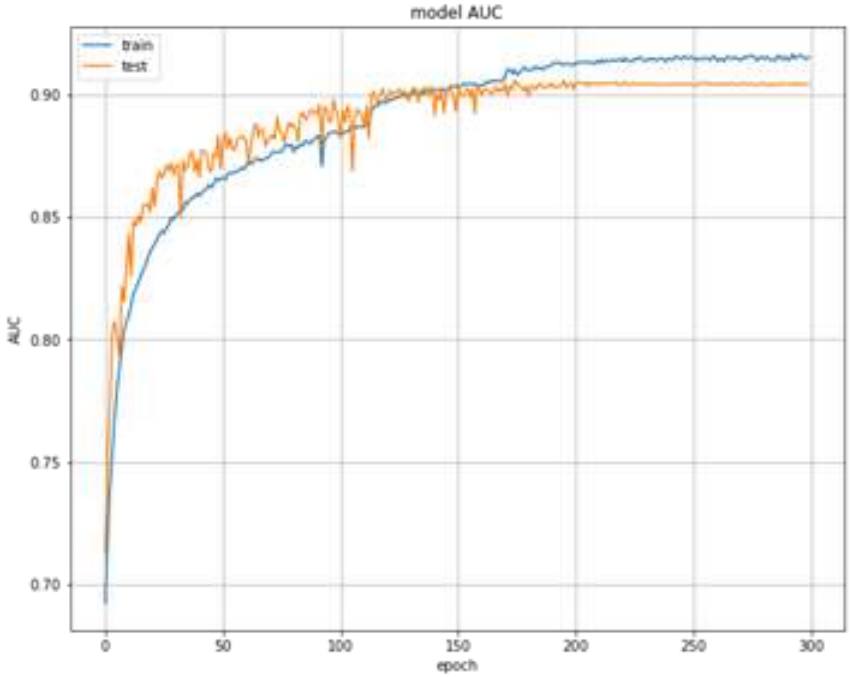

Figure 4.9 The training and validation metric was optimized over 300 training epochs

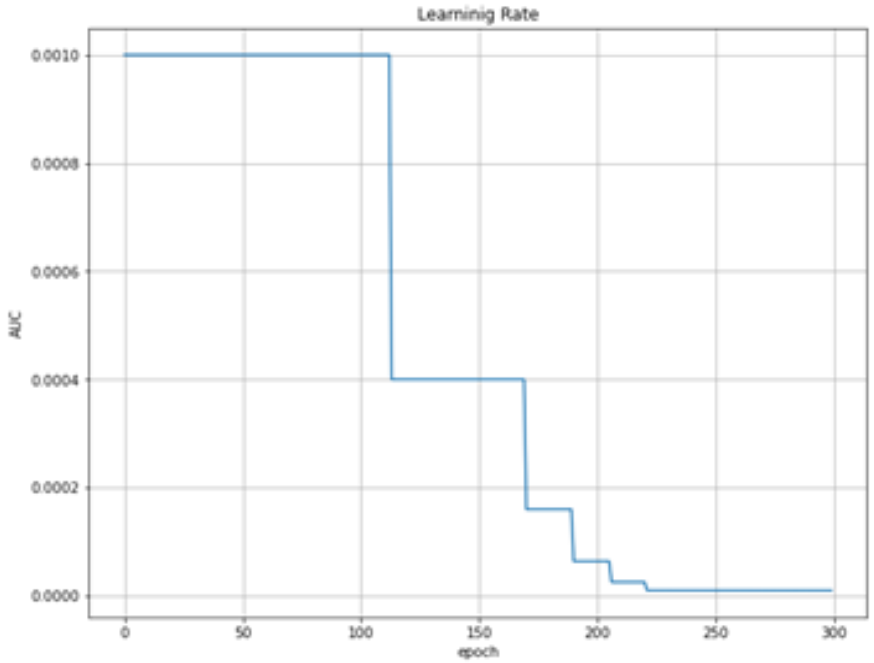

Figure 4.10 During training, the learning rate using the reduced learning rate scheduler plateaus

4.4 Training MobileNet-V3 Model

When the MobileNet-V3 model was trained, the FER2013 and RAF-DB data sets were used. The model's top layer, comprised of one thousand neurons representing one thousand classes in the ImageNet data set, is exchanged with just seven neurons that reflect our seven fundamental emotions. Suppose the validation accuracy for the executive five epochs does not increase. In that case, we utilize the scheduler learning rate to reduce the starting learning rate by a factor of 0.3, and we set the starting learning rate to 0.005.

Figures 4.11 and 4.12 illustrate the model's accuracy before and after training using the FER2013 data set. After forty epochs, we called it quits with the learning process since the model had begun to "overfit" the training data, and the validation accuracy had declined. Because of transfer learning, the model can attain an accuracy of validation that is 65.71% better than the previous best. After the first epoch, the training begins at 45% on the validation set.

The last two figures, 4.13 and 4.14, illustrate the model's accuracy and loss when trained on RAF-DB. After forty iterations, the training is terminated because the model begins to conform excessively to the data from the training set. Within the scope of this data set, the model reaches an accuracy of 83.308% during validation.

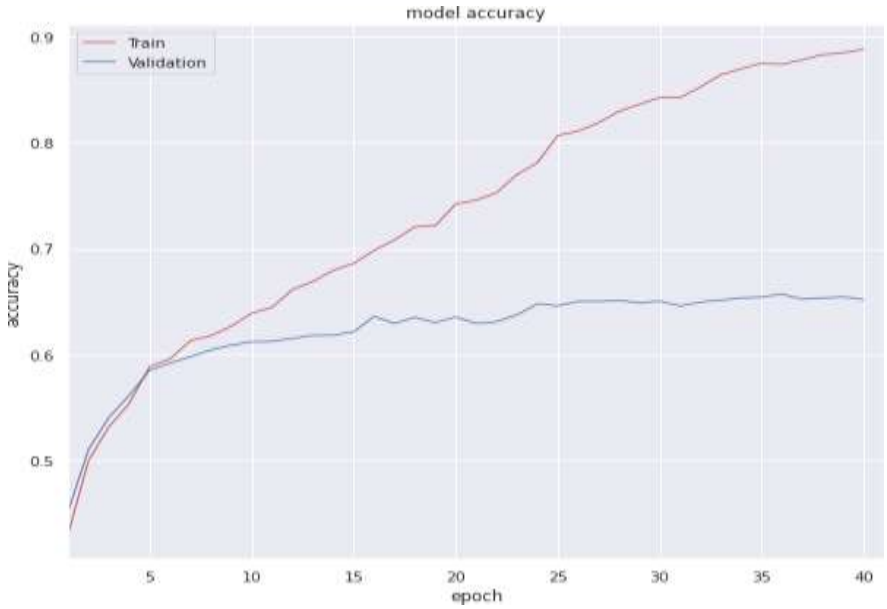

Figure 4.11 Training accuracy versus 40 training epochs on the FER2013 data set

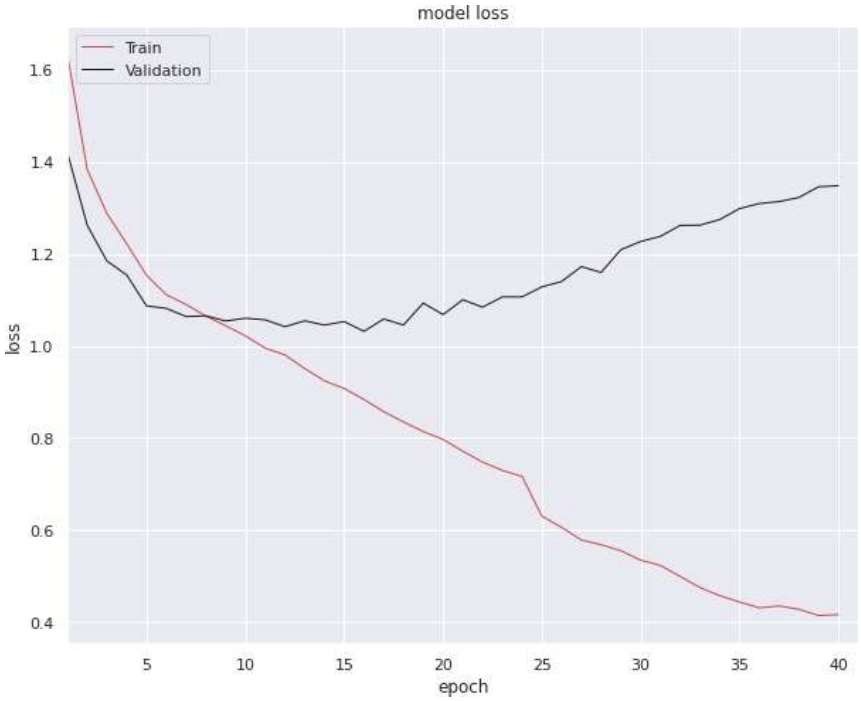

Figure 4.12 Training Cross Entropy Loss Curve of FER2013 data set

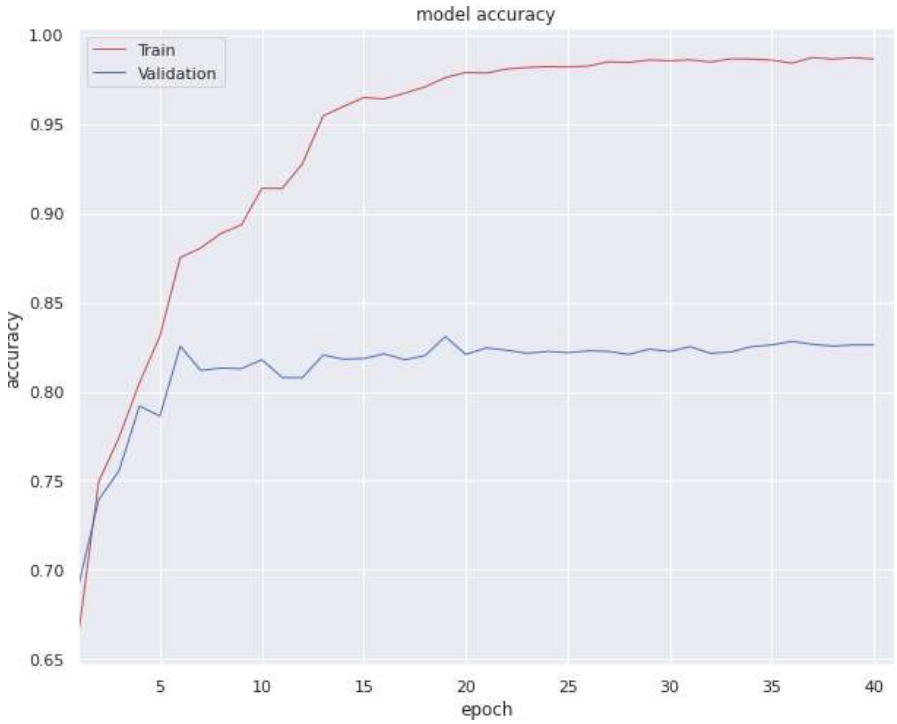

Figure 4.13 Training accuracy versus 40 training epochs on RAF-DB

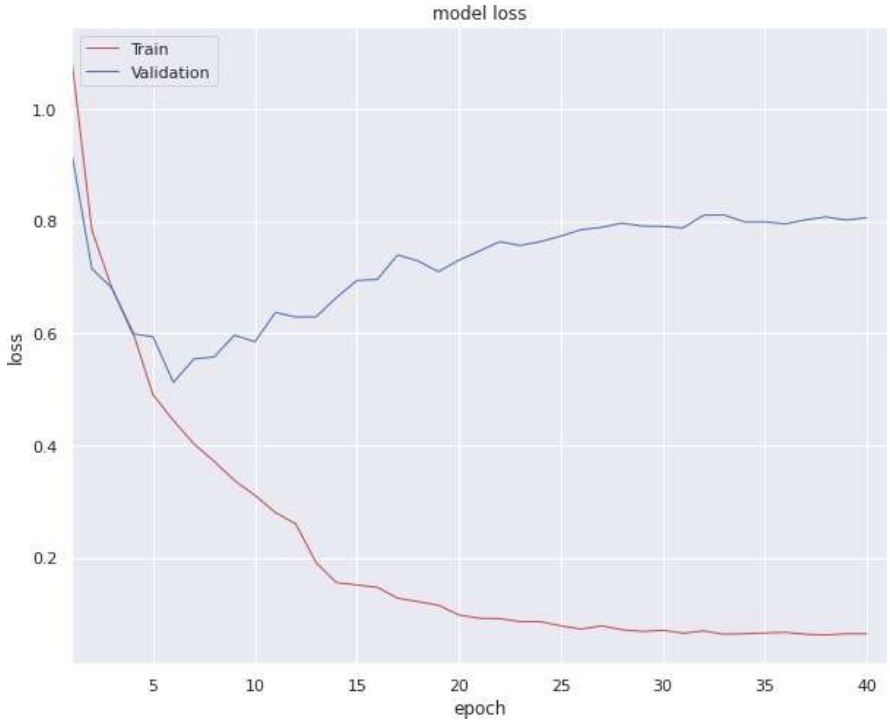

Figure 4.14 Training Cross Entropy Loss Curve on RAF-DB

4.5 Training ResNet34 Model

Transfer learning is used to train ResNet34, previously trained on the ImageNet data set. This time, it is trained on the two data sets that were picked. We decided to go with the ResNet34 version because it is suitable for the amount of information we have and because its capacity to overfit the data in the training set is suitable for the circumstances.

We change the model by exchanging the top layer, which consists of one thousand neurons, for a layer of only seven neurons. This is more suitable for our collection of classes, the seven primary feelings. We begin with a learning rate of 0.0005 for both data sets. Then, using a scheduler learning rate, we gradually decrease it until we reach saturation.

Figures 4.15 and 4.16 show the training accuracy and cross-entropy loss on the FER2013 data set. We stopped the training after 40 epochs because the model started to overfit the training set, and validation accuracy decreased. The best accuracy achieved on the validation set is 68.29%.

Figures 4.17 and 4.18 show the training accuracy and entropy loss of the Resnet 34 model on the RAF-DB. Because of the danger of overfitting, the training is terminated after 40 epochs of the algorithm. The highest possible accuracy on the validation set is 86.1%, which was attained.

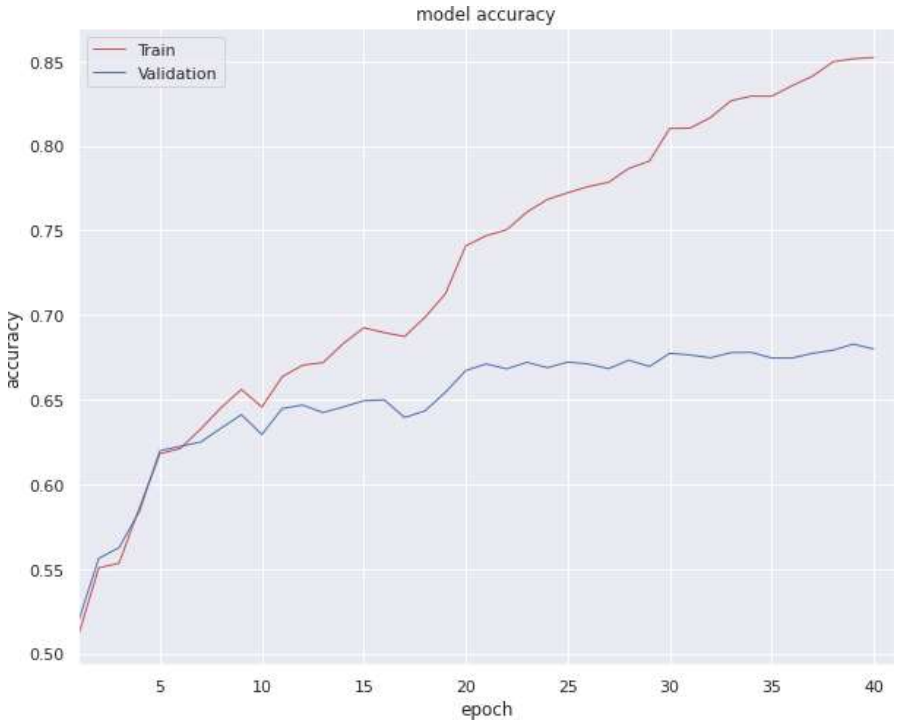

Figure 4.15 Training accuracy versus 40 training epochs on the FER2013 data set

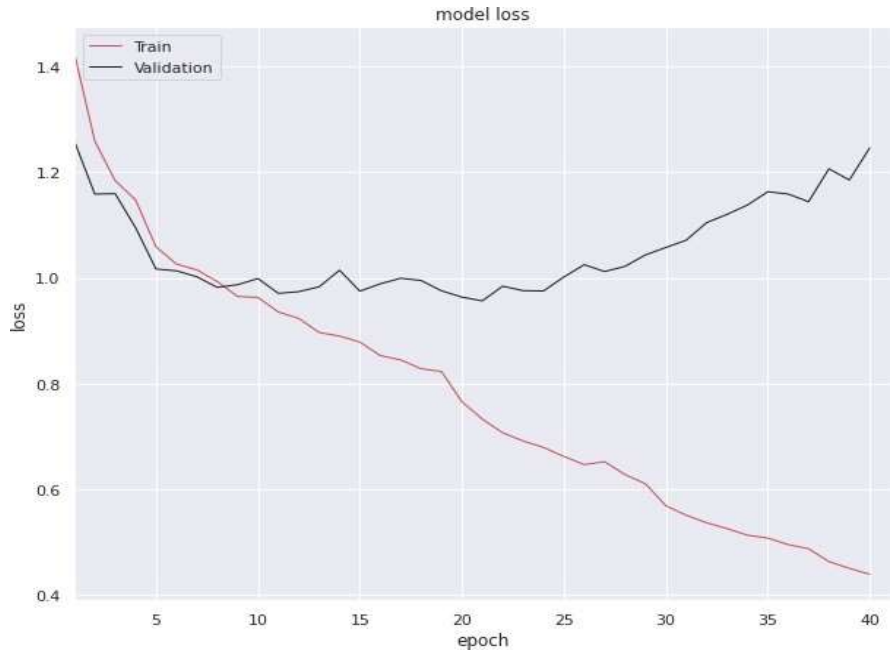

Figure 4.16 Training Cross Entropy Loss Curve of FER2013 data set

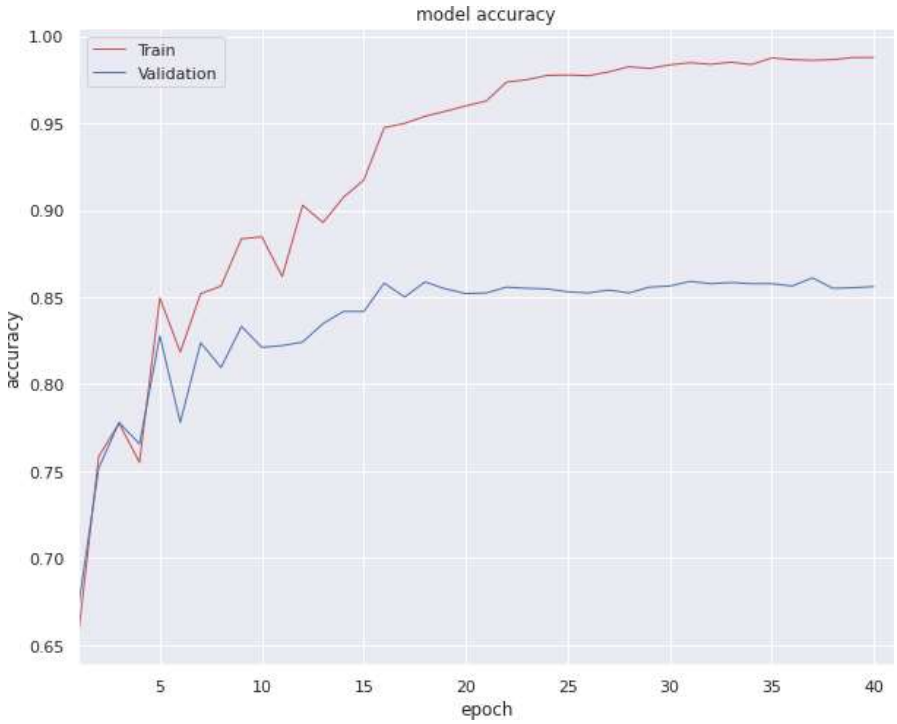

Figure 4.17 Training accuracy versus 40 training epochs on RAF-DB

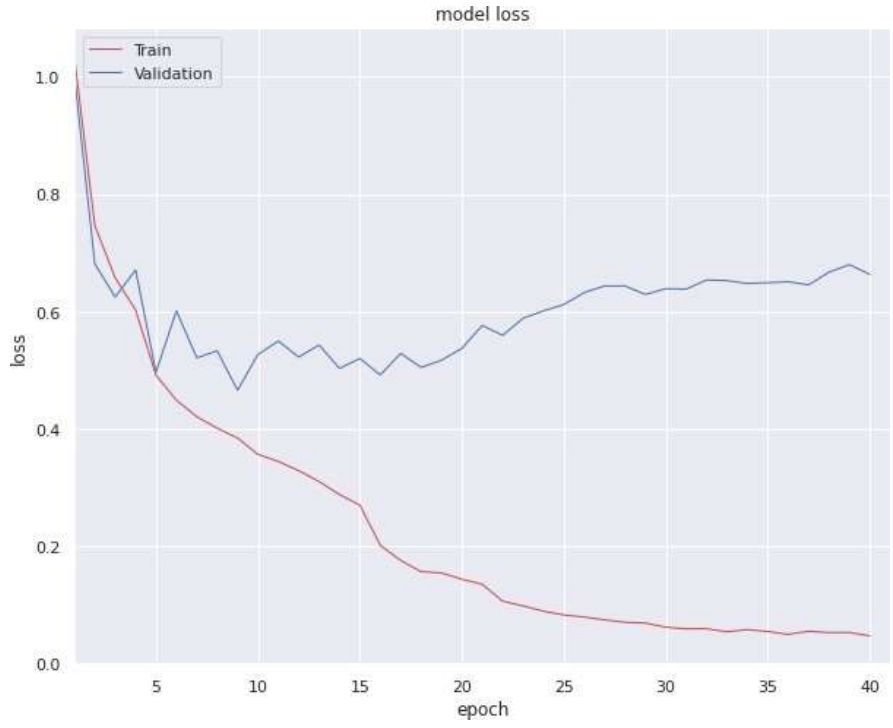

Figure 4.18 Training Cross Entropy Loss Curve on RAF-DB

4.6 Training Wider-ResNet50-2 Model

Wider-ResNet50-2 model: This "50" refers to the same number of layers as the ResNet50 model, and the "two" indicates that the width is twice as wide as ResNet50. Because this model has performed very well on various classification issues, we decided to use it to solve the current problem, hoping it would provide the same results as the other problems.

With this model, we train Wider-ResNet50-2 on both data sets mentioned above. We modify the top layer so that only seven output neurons represent the seven different classes. We employ the Adam optimizer with an initial learning rate of 0.0005 and a scheduler that reduces the learning rate when reaching a plateau.

We also use the augmentation pipeline as usual as each other algorithm we used. To set the parameters of the data augmentation module, we apply an exhaustive search between different pipelines to achieve the best pipeline that gives us the best regularization result and reduces the effect of overfitting the training data.

Figures 4.19 and 4.20 demonstrate the level of accuracy achieved after training with the Wider-ResNet50-2 custom model. After 40 epochs, The model achieved an accuracy of 69.27% on the FER2013 test; at that point, we decided to halt the learning process since the model was beginning to overfit the training data to an unacceptable degree.

Figures 4.21 and 4.22 illustrate the training and validation of the Wider-ResNet50-2 model accuracy and cross-entropy loss vs. the time spent training on RAF-DB. The model's validation accuracy was improved to an all-time high of 87.07%. Both data sets were supplemented using the same parameters for the optimizer and the learning rate scheduler and utilised the same pipeline for data augmentation.

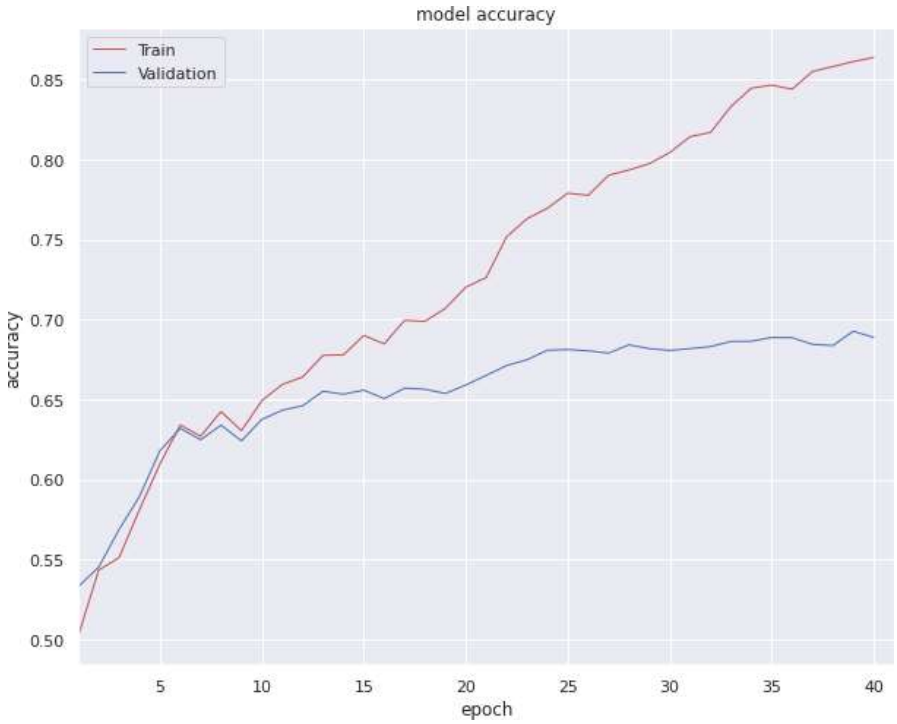

Figure 4.19 Training accuracy versus 40 training epochs on the FER2013 data set

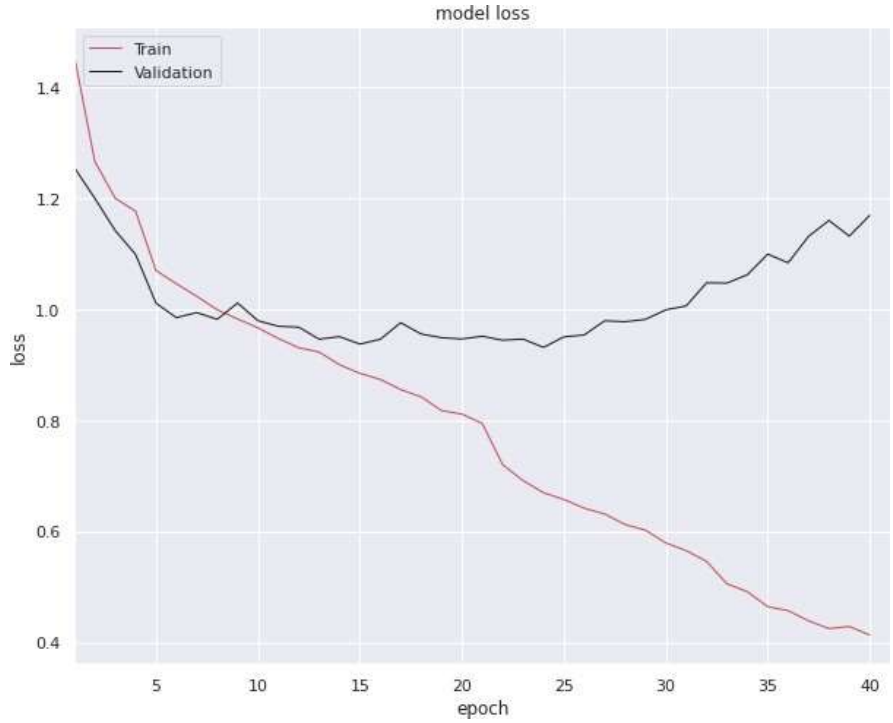

Figure 4.20 Training Cross Entropy Loss Curve of FER2013 data set

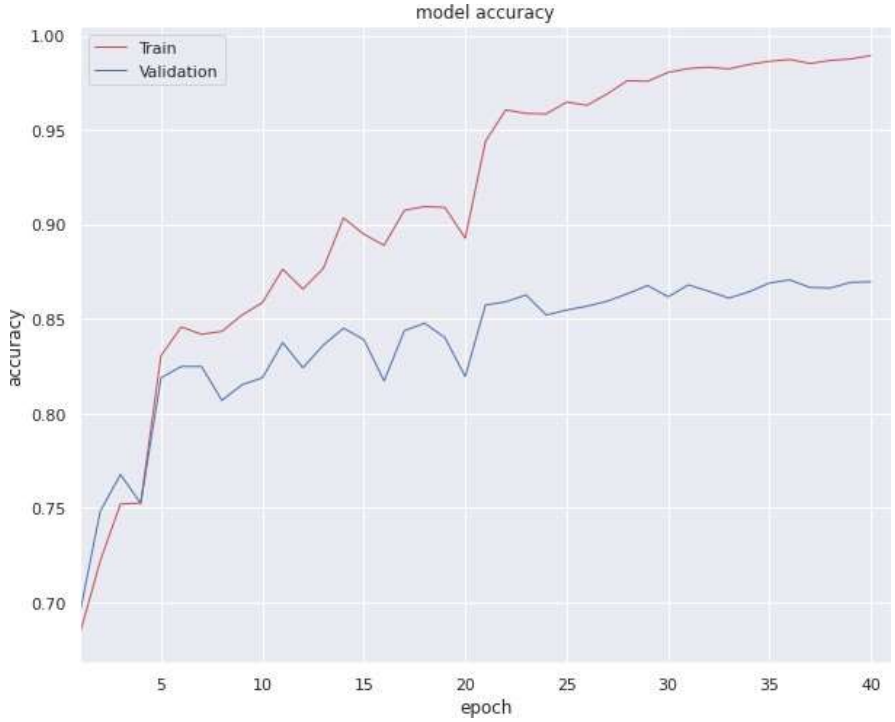

Figure 4.21 Training accuracy versus 40 training epochs on RAF-DB

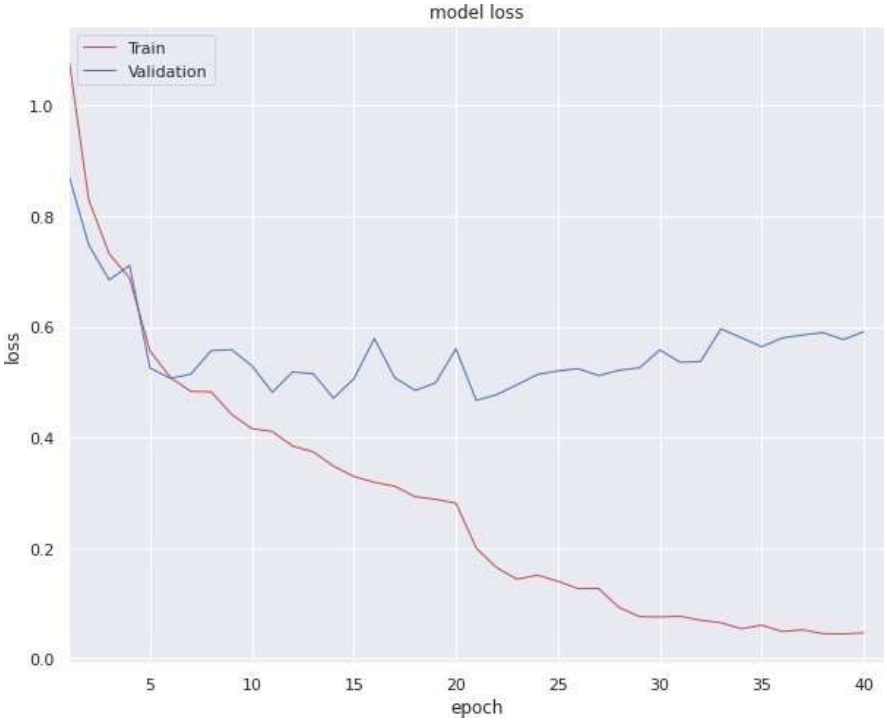

Figure 4.22 Training Cross Entropy Loss Curve on RAF-DB

4.7 Hand Based Emotion Recognition

In this section, we can determine the man's state of mind by observing the posture of his hand. Deep convolutional neural networks (CNN) are what we will be using to extract and identify the hand characteristics. The data collection, data analysis, training, and assessment of our suggested architecture are all topics that will be discussed here.

4.7.1 Data collection

The data set comprises 29 different files, each labelled with a class corresponding to one of the alphabets from the American Sign Language Akash (2018).

These three categories are very useful for classification and are also helpful for applications that involve real-time processing. In order to encourage the use of test images captured in real-world settings, the test dataset contains only 29 images.

Regarding categorization and real-time applications, these three classifications come in quite handy. In order to promote the use of real-world examples for testing, the test data collection only includes a total of 29 pictures.

Figure 4.23 The classes from A to Z to get more intuition about the data representation

However, we will employ four classes ranging from A to Z for some of our gesture recognition system's essential and significant finger movements. These classes will be derived from the data shown here. The model will go through training to learn to recognize the following four different hand gestures: the letter A (punch), the letter F (Super), the letter L (Loser), and the letter V (Victory). After that, we will instruct our model to recognize these movements and provide a suitable vocal response to each of the following as they come up.

4.7.2 Data analysis

The dataset is shown to be entirely balanced by the bar graph 4.24, and each folder has 2400 images. Continue by imagining the different images that are included in the train directory. Before studying the file sizes and the number of channels for each picture included in these directories, we will examine the first image in each subdirectory.

The following is a list of the dimensions of the images:

- The picture has a height that is equal to 200 pixels.
- The picture has a width that is equal to 200 pixels.
- There are a total of three channels available.

Similarly, we investigated the validation directory and examined the validation data set and the validation image to determine their appearance.

Within each category of our data collection, we have 600 images to represent each label. Additionally, the data for validation are entirely balanced.

Figure 4.24 Number of images for our 4 classes

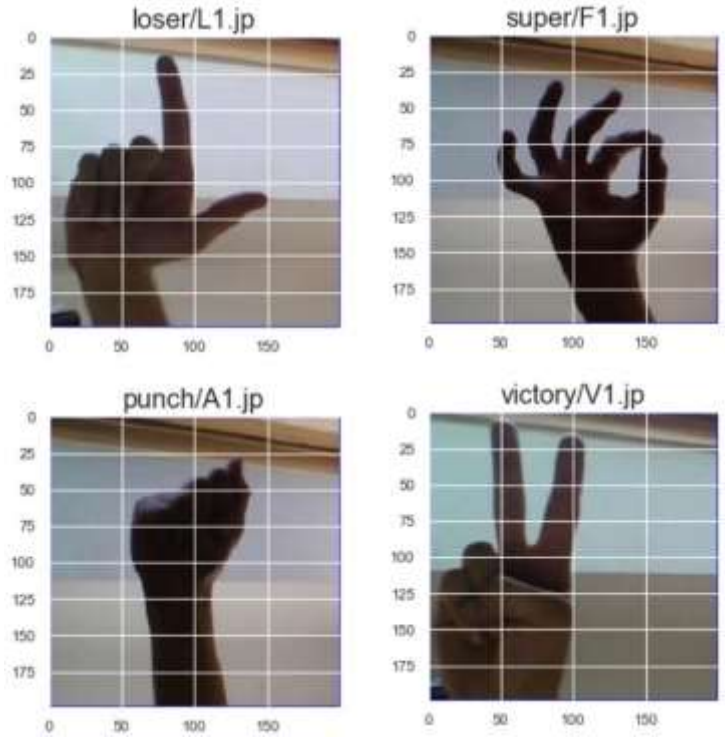

Figure 4.25 Data samples to get more intuition about the data

4.7.3 Training

Transfer learning was used for this dataset. ResNet50v2, which was previously trained on the Image Net dataset, was used. Data augmentation, as discussed in the FER2013 dataset, was also used.

This is our parameter of the date augmentation generator:

- The Random Rotation is set to = 30 deg
- The Shear Range is set to = 0.3
- The Zoom Range is set to = 0.3
- The Width Shift Range is set to = 0.4
- The Height Shift Range is set to = 0.4
- The Horizontal Flip is set to = True

The model underwent 50 training epochs, utilizing an initial learning rate of 0.001. We employ the Adam optimizer for training. Also, we Reduce the Learning rate using the Reduce on Plateaulearning rate scheduler.

We achieved an accuracy of 99.91% on the validation data set. This approach is very accurateconcerning the previous work published on this data set.

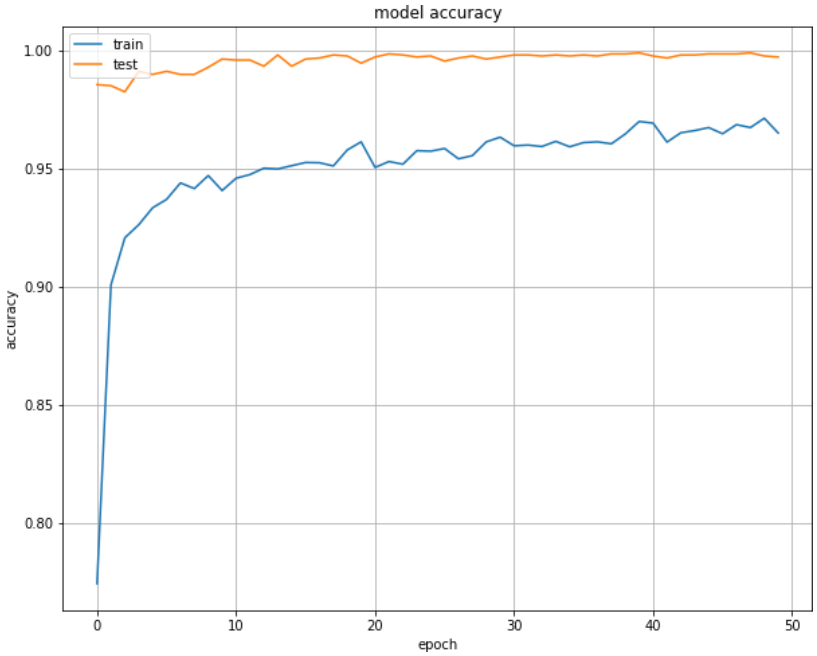

Figure 4.26 Training accuracy on training and validation data set

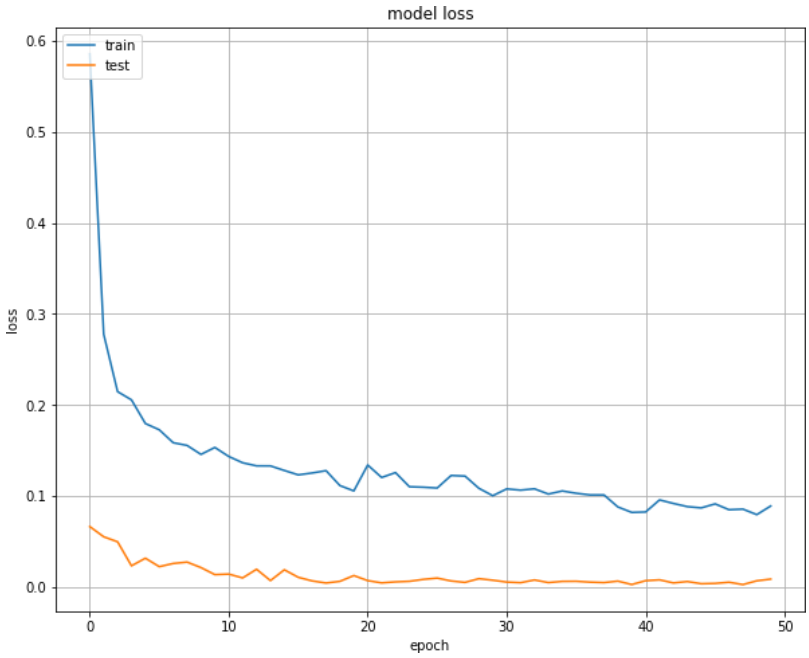

Figure 4.27 Training loss on both training and validation set

4.8 Conclusion

This part illustrates the two data sets we used and samples some examples to get more intuition about each data set we use. Also, we discussed each model training section and illustrated the training graph of each one of them. We calculate the validation accuracy and loss to get a more accurate intuition of the performance of the models on unseen data.

The Wider-ResNet50-2 gets the highest validation score on the FER2013 data set and RAF-DB. It gets a 69.27% and 87.07% validation accuracy on FER2013 and RAF-DB, respectively.

The following section evaluates each model on the test set and additional assessment metrics, including F1-score, precision, and recall. In order to comprehend the many kinds of mistakes our models make, we also compute the confusion matrix.

5. RESULT & DISCUSSIONS

5.1 Introduction

The duration of the training period was undergone, together with the precision of the training and validation sets. We will discuss this in the last chapter, "Training and Experiments". In this part, we will discuss the performance of each model on its own, and we will also acquire a deeper understanding of how our models operate when applied to unknown data. On the other hand, we evaluate our models by applying the test set presented in this section.

This section provides a concise overview of the model's efficiency by leveraging the Confusion Matrix and additional measurements like accuracy for each label, recall, and the F1 score. Let us begin with the confusion matrix, which illustrates the myriad of mistakes produced by the model and the frequency with which each error occurs along the different labels. It is a method for condensing the effectiveness of a classification system. Classification accuracy may be deceiving if the data set has more than two classes or the number of observations varies between classes.

By computing a confusion matrix, Precision, Recall, and F1-Score, we may better understand the successes and failures associated with the classification model. Instead of focusing on the classification system's overall accuracy, these aspects are concerned with developing a more nuanced understanding of the system's performance.

Precision pertains to the exactness of positive predictions. In contrast, recall pertains to capturing all positive events, and the F1 score provides a balanced measure that considers both precision and recall. The harmonic mean is used in the F1 score to give equal weight to precision and recall.

5.2 Baseline Model Evaluation

During the rounds of review and development, we concluded that we should, from the very beginning, build a starting point model for our work direction. Our objective is to study the effect of transfer learning on all different learning models. Figure 5.1 can analyse the confusion matrix.

In addition, the first model's accuracy, recall, and score for each label were computed, and the results are provided in Table 1. Through this study, we can understand how the model works on data that has not yet been examined and determine which labels have a more significant number of mistakes.

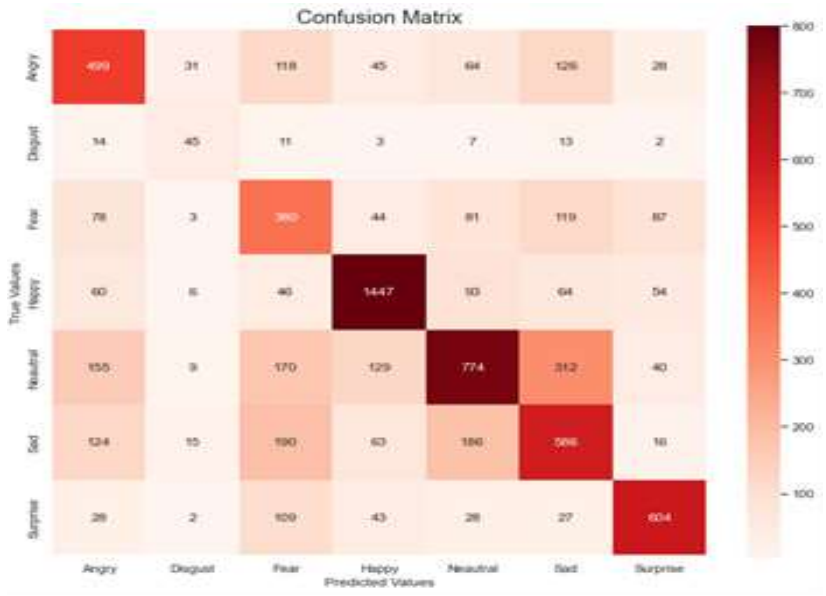

Figure 5.1 The first CNN model for our confusion matrix

It allows us to understand the importance of taking advantage of the problem's difficulty and emphasizes the benefit of transfer learning. It is necessary to test the dataset with models that can hold the data and use it effectively.

Table 5.1 The baseline model's classification report displays the F1 score, recall, and precision for each label

Label	Precision	Recall	F1-Score	Num. of images
Angry	0.56	0.50	0.53	958
Disgust	0.48	0.41	0.44	111
Fear	0.47	0.34	0.39	1024
Happy	0.80	0.83	0.82	1774
Neutral	0.51	0.66	0.57	1233
Sad	0.50	0.48	0.49	1247
Surprise	0.72	0.73	0.72	831
Accuracy	–	–	60.85 %	7178

5.3 MobileNet-V3

We determine the MobileNet-V3 model's accuracy, recall, and F1-score for every label to grasp better how it performed on the RAF-DB and FER2013 datasets. One way to normalize the confusion matrix is to divide it by the count of images that correspond to each label. This exercise aims to evaluate the dispersion of the percentage of mistakes, which is depicted quite similarly in Figure 5.2.

The categorization report for the model may be viewed in Table 2. The table provides calculations of the f1-score, degree of accuracy, and level of recall for each label linked with one of the seven primary emotions. In addition, the fact that such a massive number of images were used in the research offers further context for how our MobileNet-V3 model performed effectively on FER2013. A score of 0.79 on the F1 scale is awarded to the surprise label.

Figure 5.3 shows the normalized and normalized confusion matrices for the effective network model for the RAF-DB test set. Both of these matrices are available for viewing. Compared with other databases, RAF-DB has a more potent error distribution. This error distribution is because a colour image with a pixel size of 100 by 100 is present in the data. This situation developed because the image allows the model to get additional data about the data.

The report on categorizing the model inside the RAF-DB can be seen in Table 3. Based on the criteria of the test set, it has been found that the model's accuracy is 83.447%. A high number of positive examples in the training data allows the model to predict this class accurately, so it has a score of 0.93 for correctly predicting photos of smiling people. Worst of all, the F1 score measure threshold for surprise, sadness, and neutrality is more than 0.8.

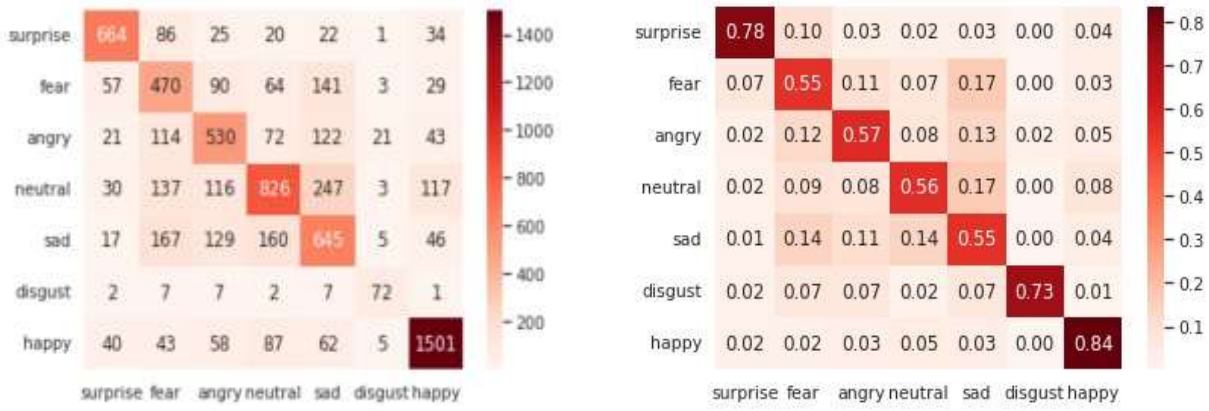

Figure 5.2 FER 2013's test set contains two matrices: one for confusion and one for normalized confusion

Table 5.2 The MobileNet-V3 model assessment report On the FER 2013 test set, showing the recall, precision, and F1 score for every label

Label	Precision	Recall	F1-Score	Num. of images
Surprise	0.78	0.80	0.79	831
Fear	0.55	0.46	0.50	1024
Neutral	0.56	0.67	0.61	1231
Angry	0.57	0.55	0.56	955
Disgust	0.73	0,65	0.69	110
Sad	0.55	0.52	0.53	1246
Happy	0.84	0.85	0.84	1771
Accuracy	–	–	66.09%	7178

Figure 5.3 The RAF-DB's test set contains two matrices: one for confusion and one for normalized confusion

Table 5.3 The MobileNet-V3 model's evaluation report On the RAF-DB test set, showing the recall, precision, and F1 score for every label

Label	Precision	Recall	F1-Score	Num. of images
Surprise	0.80	0.84	0.82	323
Fear	0.68	0.59	0.63	74
Disgust	0.65	0.39	0.49	160
Happy	0.92	0.94	0.93	1159
Sad	0.,78	0.83	0.81	467
Angry	0.73	0.73	0.73	160
Neutral	0.78	0.79	0.79	665
Accuracy	–	–	83.44%	3008

5.4 ResNet34 Model Evaluation

The results of the ResNet34 models are described in a condensed manner. This section assesses our model's effectiveness on the test dataset by calculating metrics that provide insights into its overall performance. Specifically, we examine the confusion matrix. A normalized confusion matrix was generated for each dataset. Additionally, we provide a categorization report that includes the F1 score, accuracy, and recall for each label corresponding with the seven emotions.

The above confusion matrix is shown in Figure 5.4. Within the FER2013 dataset, The ResNet34 model's confusion matrix was normalized. The accurate prediction of surprise labels may be achieved with 86% and 78% precision, respectively, using this approach. In addition, it demonstrates progress in reducing errors in naming, which has achieved 75% accuracy. However, regarding fear labels, error rates have been reported where the model appears to be confused between these two sets.

The classification report in Table 4 was produced using the dataset in conjunction with the MobileNet V3 model comparison. Compared to the MobileNet V3 Model, it is possible to conclude that each label obtains a score of F1 on the test set. This obtained label is based on the model's accuracy, which is 68.57 per cent.

When looking at Figure 5.5, the confusion matrix is shown in the last two figures. When applied to the RAF DB test set, this is the normalised confusion matrix for the ResNet34 model. Based on these results, it is not difficult to see that the model works well on this dataset, obtaining a degree of accuracy that is satisfactory to researchers.

Information regarding our models' performance on the RAF DB test set is provided in Table 5. To be more precise, they show F1-score, accuracy, and recall for each of the seven emotions we go through. The model's accuracy in testing was 86.32 per cent. She also obtained a score of 0.94 for naming and F1 scores and above 0.80 for three additional names: surprised naming, sad naming, and neutral naming.

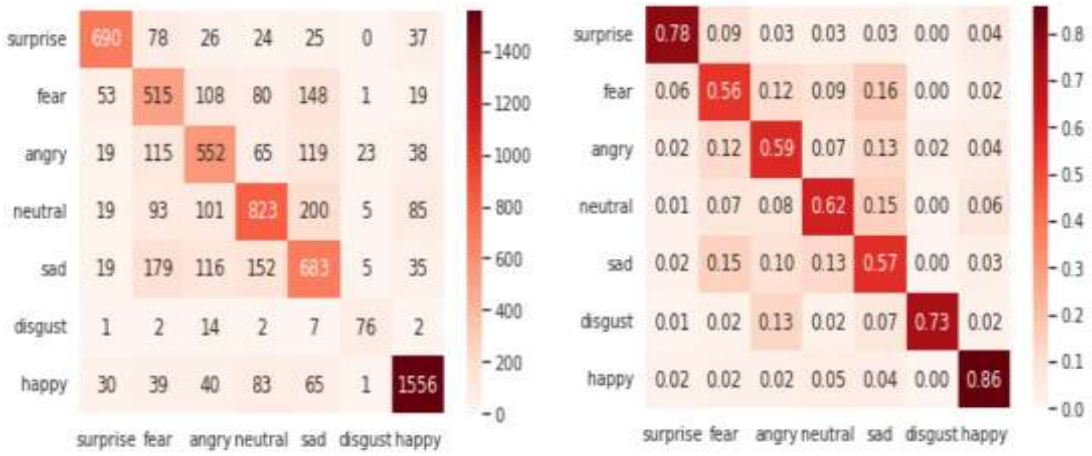

Figure 5.4 Each label's precision, recall, and f1 score are included in the classification report of the ResNet34 model in the FER2013 test set

Table 5.4 Every label's precision, recall, and F1 score are shown in the classification report for the ResNet34 model on the FER2013 test set

Label	Precision	Recall	F1-Score	Num. of images
Surprise	0.78	0.83	0.81	831
Angry	0.59	0.58	0.58	957
Fear	0.56	0.50	0.53	1021
Sad	0.57	0.55	0.56	1247
Neutral	0.62	0.67	0.64	1229
Happy	0.86	0.88	0.87	1772
Disgust	0,73	0.68	0.71	111
Curacy	–	–	68.57%	7178

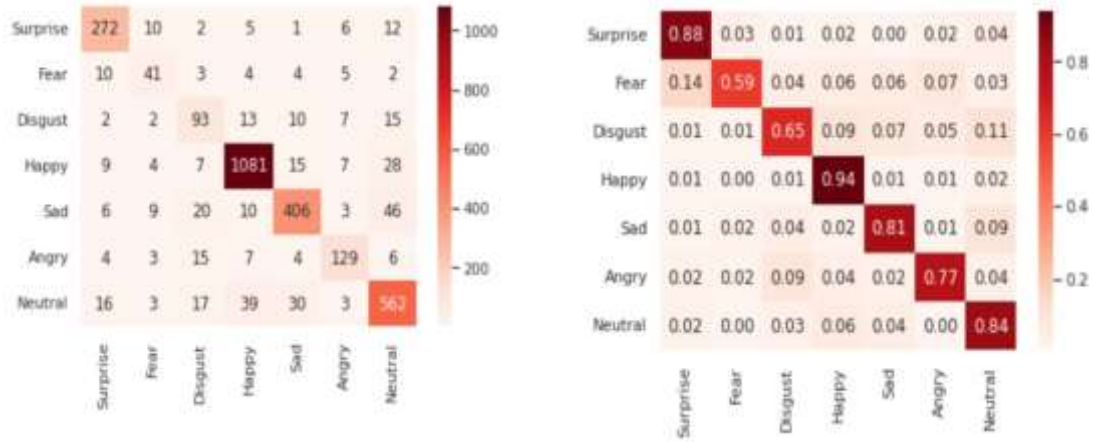

Figure 5.5 The RAF-DB test set's Normalized Confusion Matrix and Confusion Matrix, as calculated and represented

Table 5.5 The classification report of the ResNet34 model on the RAF-DB test set presents the precision, recall, and F1 score for each label

Label	Precision	Recall	F1-Score	Num. of images
Surprise	0.88	0.85	0.87	319
Disgust	0.65	0.59	0.62	157
Fear	0.59	0.57	0.58	72
Angry	0.77	0.81	0.79	160
Sad	0.81	0.86	0.84	470
Happy	0.94	0.93	0.94	1159
Neutral	0.84	0.84	0.84	671
Accuracy	–	–	86.32%	3008

5.5 Wider-ResNet50-2 Model Evaluation

We evaluate the performance of the Wider ResNet50 2 model in the context of the test set. An idea of how well our model works on data that has not yet been seen can be obtained by analyzing the normalized confusion matrix and the confusion matrix. When applied to the FER2013 dataset, the Wider-ResNet50-2 model provides the largest possible result of 69.41%, with test accuracy. Compared to other models, the original broader ResNet50-2 model contains 69 million parameters, which is a very large amount. In addition, the complexity of this model is very high. Although the number of parameters has been reduced to 66 million due to our customization, this amount is still excessive compared to other architectures we use.

In Figure 5.6, both the normalized and normalized confusion matrices for the model. The model was applied to the FER2013 validation set. Also, the Wider-ResNet50-2 model exhibited a decrease in error across all classes compared to the other models. This error decrease may be demonstrated by contrasting the main diagonal elements of the confusion matrix for the Wider-ResNet50-2 model with the main diagonal elements of the confusion matrices for the other models.

The labels' precision, recall, and F1-score metrics are displayed in Table 6, along with the count of images linked to each label that goes into calculating these metrics. The

data clearly shows that the F1 score for every label has improved. The test set also shows a very high accuracy, at 69.41%.

The normalized confusion matrix and confusion matrix for the Wider-ResNet50-2 model applied to RAF-DB are shown in Figure 5.7. The Wider-ResNet50-2 model outperforms all other models on this dataset, with an accuracy rating of 87.23% on the test set.

In order to provide a more comprehensive comprehension of the functioning of the Wider-ResNet50-2 model on unseen data, Table 7 presents further details regarding the model's performance on each label.

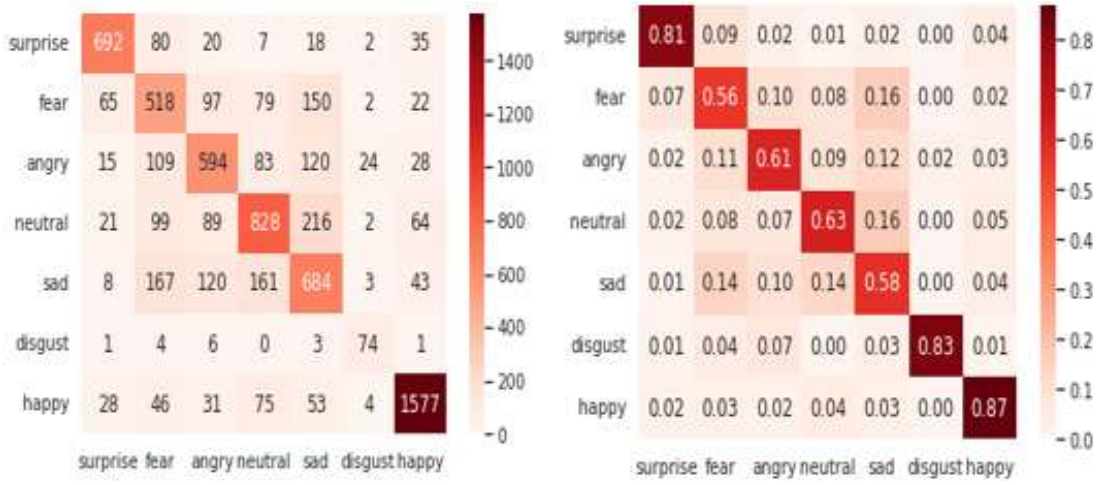

Figure 5.6 The FER 2013 test set's Normalized Confusion Matrix and Confusion Matrix, as calculated and represented

Table 5.6 The classification report of the Wider ResNet50 -2 model on the FER2013 test set presents the precision, recall, and F1 score for each label.\

Label	Precision	Recall	F1-Score	Num. of images
Surprise	0.81	0.83	0.82	830
Angry	0.61	0.62	0.62	957
Fear	0.56	0.51	0.53	1023
Sad	0.58	0.55	0.56	1244
Neutral	0.63	0.67	0.65	1233
Happy	0.87	0.89	0.88	1770
Disgust	0.83	0.67	0.74	111
Accuracy	–	–	69.41%	7178

Figure 5.7 The RAF-DB's test set's Normalized Confusion Matrix and Confusion Matrix, as calculated and represented

Table 5.7 Classification report of the Wider-ResNet50-2 model on RAF-DB test set shows each label's precision, recall, and F1 score

Label	Precision	Recall	F1-Score	Num. of images
Surprise	0.88	0.85	0.86	327
Disgust	0.68	0.56	0.61	153
Fear	0.69	0.60	0.64	70
Sad	0.84	0.87	0.86	466
Happy	0.94	0.94	0.94	1163
Neutral	0.84	0.88	0.86	669
Angry	0.80	0.81	0.80	160
Accuracy	–	–	87.23%	3008

5.6 Comparisons model

In this part of the article, we will provide the models' results and evaluate them according to the degree of accuracy with which they completed the test. We use a total of four different models to analyse the data set that was collected in FER2013. In order to get started, we will be using the basic model built from the ground up. Following that, we employ three distinct Deep Convolutional neural network architectures: MobileNet, ResNet34, and Wider Resent. The second data set, referred to as RAF-DB, is the only one to which we restrict these deep learning methods since the accuracy of these techniques is more promising.

Table 8 displays the results of the several models run on the FER2013 dataset; as can be seen, the deep learning model outperformed the baseline model in terms of accuracy. The most efficient model is highlighted in grey in the accompanying table. With a rate of 69.41%, the ResNet50-2 model suggested by Zagoruyko and Komodakis (2016) performs best in testing this dataset. It is critical to consider the complexity and number of model parameters while implementing a suitable model.

Table 9 also shows all of the results from RAF-DB. Our analysis of the FER2013 results led us to conclude that the three deep learning models presented here have great potential for solving the emotion recognition problem. Hence, we restrict our use of these DL models to the RAF-DB dataset. Specifically, the ResNet50-2 model proposed by Zagoruyko and Komodakis (2016) outperforms the competition on this dataset, achieving an impressive test accuracy of 87.23%.

Table 5.8 Results presented for models evaluated on the FER2013 test set

Model	Test Accuracy	Number On Parameters
Base Line	60.85%	1.67M
MobileNet-V3	66.09%	4.21M
ResNet34	68.57%	21.29M
Wider-ResNet50-2	69.41%	66.48M

Table 5.9 Results presented for models evaluated on the RAF-DB test set

Model	Test Accuracy	Number On Parameters
MobileNet-V3	83.447%	4.21M
ResNet18	86.32%	21.29M
Wider-ResNet50-2	87.23%	66.48M

Figure 5.8 Random images from the test set with its prediction (Wider-ResNet50-2) and real labels

5.7 Robustness of Our Approach

In order to ascertain the robustness and dependability of our methodology, we implemented our pipeline on two distinct data sets. Furthermore, we do not rely on a single algorithm to tackle the issue. Conversely, we investigate several methodologies and frameworks to construct robust models that precisely depict our research.

We extensively search to determine our model's optimal hyperparameters and augmentation processes. The aim is to have ultra-robust parameters that provide the best results reflecting each model.

To ensure the robustness of our technique, we address all of these concerns and apply our process to two publicly accessible datasets. When we compare our findings with those of other researchers, we can get a more accurate model and attain the best possible outcomes on the data sets in question.

Tables 10, 11, and 12 compare our best findings with other state-of-the-art results on both data sets, the RAF-DB and the FER2013. This research model, Wider-ResNet50-2, beats all other research models regarding its accuracy on FER2013 and RAF-DB. At the same time, as some researchers rely on several datasets to combine them into a single larger dataset, we rely only on each dataset on its own, without using any merging technique. Our use of customisation layers for each model reduces the total number of parameters for the models while simultaneously increasing the amount of time required for inference compared to the initial models.

Table 5.10 Comparison of FER2013 dataset of the proposed approach and another state-of-the-art result

Models	Test Accuracy
Feng and Ren Suryanarayana et al., (2021)	66.67%
Devries et al. Devries et al., (2014)	67.21%
Mollahosseini et al.,Hussain andAl Balushi (2020)	69.30%
Ours (Wider-ResNet50-2)	**69.41%**

Table 5.11 Comparison of the proposed approach and another state-of-the-art result on the RAF dataset

Models	Test Accuracy
gACNN Li et al., (2019)	85.07%
DLP-CNN Li et al., (2017)	84.13%
Ours (Wider-ResNet50-2)	**87.23**%

5.8 Evaluation of our Pipeline Using Face Detection Technique (MTCNN model)

This section will publish our performance model, Face Detection Model (MTCNN). To start, we will use an image from the website in Figure 5.9. In the processing stage, we include an image enhancement layer that improves the image's contrast. While this enhancement layer may have no effect, it does improve overall image quality since laptop camera images tend to be grainy.

Figure 5.10 displays the prior image, following the application of the enhancement layer and the face detection model. The final illustration shows the output image after making predictions, highlighting the bounding boxes around the detected faces generated by the MTCNN model and the emotional predictions generated by Wider ResNet50 2. Additionally, one step is performed before making the predictions, known as convolution.

The facial picture was placed into grayscale and scaled to 48x48 pixels to ensure it was compatible with the FER2013 Kaggle data set.

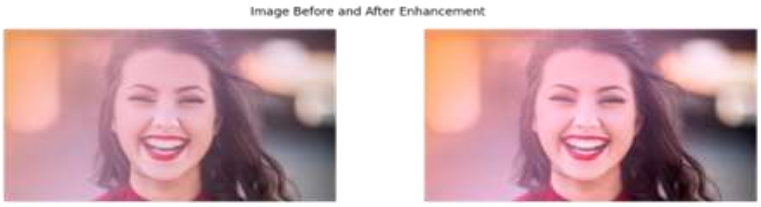

Figure 5.9 Test image before and after the improvement procedure

Figure 5.10 The face image was extracted using the MTCNN face detection model

Figure 5.11 Pipeline Output Image

5.9 Conclusion

Here, we presented the results of each model separately and examined them in greater detail using several metrics functions. We first identify the advantages and disadvantages of each model on its own, then we compare them all to one another to provide a sufficient description and comparison.

With the help of ResNet34 He et al., (2016) and MobileNet V3Howard et al., (2017), we could show that advanced convolutional neural network designs can perform well. Zagoruyko and Komodakis (2016) have shown that wider ResNet50 2 can potentially provide outcomes in emotion recognition. It is essential to recognise the potential for technological improvements that they possess.

The performance of the models was assessed on the test set by generating the confusion matrix and the normalized confusion matrix. These matrices allow us to understand how our models perform with the given data. It allows to understand better how our models perform with data. With a test accuracy of 69.41%, the ResNet50-2 Zagoruyko and Komodakis (2016) model is the one that yields the best results when applied to the FER2013 data set. Since it has 66 million parameters, the original ResNet50-2 model developed by Zagoruyko and Komodakis (2016) is excessively complicated and extensive compared to existing models.

On RAF-DB, the wider ResNet50-2 Zagoruyko and Komodakis (2016) get 87.23% test accuracy, achieving the best result on this data set. Also, we deploy our pipeline with the MTCNN face detection model and test it with real-life data to determine how the model performs in real-life data and get more intuition about its behaviour.

6. CONCLUSION AND FURTHER DIRECTIONS

6.1 Outline of The Contribution

Studying facial expressions, which communicate much nonverbal information, may improve understanding of human feelings and social interactions. The study's approach was fruitful in identifying facial expressions presenting an original point of view on issues previously covered in the relevant body of research. With our cutting-edge approach, it can now distinguish and classify facial emotions more accurately and effectively, which decreases computing costs and demands while boosting picture recognition rates. The face picture classification procedure was made better with the help of the constructed model. According to the findings of our study, the efficiency of facial expression recognition techniques may be improved using deep learning to achieve higher levels of accuracy, improved face recognition, and more accurate interpretation of facial characteristics and emotions.

Identifying emotions through expressions is an exciting subject with applications in various fields, including safety, health and human-machine interactions. Researchers in this field are focused on enhancing computer predictions by developing techniques to interpret, categorize, and extract emotions from expressions. The remarkable success of learning has led to the adoption of diverse approaches aimed at improving performance.

The limits of our study were examined in this area, and how these constraints might be considered for improvement.

6.2 Limitation

By integrating natural language processing (NLP) with automatic facial expression recognition, it is feasible to augment the complexity of automatic facial expression recognition systems. If this future advancement is implemented, it has the potential to impact significantly the e-health system and the delivery of healthcare services.

The combination of language and facial expressions for predicting human emotions would be a significant improvement. This integration would provide the system with information enabling more precise emotion prediction.

In our approach, we primarily focus on expressions. Give less emphasis to hand positions. However, it is essential to note that our work has limitations as we only address the issue using Face emotion datasets FER2013 and RAF DB. Additionally, we employ CNN architecture to extract information related to facial expressions.

6.3 Overal Conclusion

Using a CNN, we dug into the matter. A study was conducted to determine the impact of the learning blocks on the model's performance using the FER2013 and RAF DB datasets. A foundational model was developed based on our prior experience utilizing Convolutional Neural Networks (CNNs) on datasets by determining the appropriate number of layers and neurons for each layer. The learning rate and other augmentation parameters were fine-tuned through trial and error to achieve results.

Additionally, another technique was explored by leveraging ResNet50v2, which had been pre-trained on the ImageNet dataset. This leverage allowed us to predict four labels based on hand position: punch, Super, Loser and Victory.

With the possibility of merging our model with others, face recognition systems using the FER2013 data set could achieve even higher accuracy. This achievement is because it obtains a respectable degree of accuracy while reducing the required processing work. Despite the FER2013 data set's complexity and low sample size per class, it is possible to increase accuracy by appropriately increasing the number of samples in each class. The small sample size of the FER2013 data set is to blame for this.

We employ transfer learning approaches to re-train ResNet34 (He et al., 2016), MobileNet-V3 (Howard et al., 2017), and Wider ResNet50-2 (Zagoruyko and Komodakis 2016) using the open-source data sets FER2013 and RAF-DB in order to

overcome this issue. Utilizing deep CNN architectures that have been pre-trained demonstrated excellent performance on these data sets.

We test our models on the two different data sets and construct the confusion matrix as well as the normalized confusion matrix so that we can have a better idea of how our model reacts when it is presented with data that it has not seen before.

Based on the FER2013 data set, the Wider ResNet50-2 (Zagoruyko and Komodakis 2016) model has the highest level of performance, with a test accuracy of 69.41%. Compared to other models, the original Wider ResNet50-2 model contains 68 million parameters, making it excessively difficult and enormous. With our customisation, however, we can reduce the total number of parameters to 66 million.

On RAF-DB, the wider ResNet50-2 (Zagoruyko and Komodakis 2016) gets 87.23% test accuracy, achieving the best result on this data set. Also, we deploy our pipeline with the MTCNN face detection model and test it with real-life data to determine how the model performs in real-life data and get more intuition about its behaviour.

REFERENCES

Afriyie, R., Asante, M., and Onyema, E. M. 2020. Implementing morpheme-based compression security mechanism in distributed systems. *International Journal of Innovative Research & Development (IJIRD)*, 9(2); 157–162.

Ahirwar, M. K., Shukla, P. K., and Singhai, R. 2021. Cbo-ie: a data mining approach for healthcare iot dataset using chaotic biogeography-based optimization and informationentropy. *Scientific Programming*, 2021.

Akash, M. "Image data set for alphabets in the American Sign Language." https://www. kaggle. com/grassknoted/asl-alphabet. Accessed on May 2 (2018): 2021.

Allen & Pease, B. "The definitive book of body language." (2004).

Bhatt, R., Maheshwary, P., Shukla, P., Shukla, P., Shrivastava, M., and Changlani, S. 2020. Implementation of fruit fly optimization algorithm (ffoa) to escalate the attacking efficiency of node capture attack in wireless sensor networks (wsn). Computer Communications, 149; 134–145.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*; 4690–4699.

Devries, T., Biswaranjan, K., and Taylor, G. W. 2014. Multi-task learning of facial landmarks and expression. In *2014 Canadian conference on computer and robot vision*; 98–103. IEEE.

Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., and Gedeon, T. 2017. From individual to group-level emotion recognition: Emotiw 5.0. In *Proceedings of the 19thACM international conference on multimodal interaction*; 524–528.

Efron, David. "Gesture, race and culture: A tentative study of the spatio-temporal and" linguistic" aspects of the gestural behavior of eastern Jews and southern Italians in New York City, living under similar as well as different environmental conditions." (1972).

Ekman, P., Friesen, W. V., O'sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K.,Krause, R., LeCompte, W. A., Pitcairn, T., Ricci-Bitti, P. E., et al., 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4); 712.

Gunes, H. and Piccardi, M. 2005. Fusing face and body gesture for machine recognition of emotions. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, pages 306–311. IEEE.

Gunes, H. and Piccardi, M. (2006). A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *18th International conference on pattern recognition (ICPR'06)*, volume 1, pages 1148–1153. IEEE.

Gunes, H., Shan, C., Chen, S., and Tian, Y. (2015). Bodily expression for automatic

affect recognition. *Emotion recognition: A pattern analysis approach*, pages 343–377.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Heravi, E. J., Aghdam, H. H., and Puig, D. 2016. Classification of foods using spatial pyramid convolutional neural network. In *CCIA*; 163–168.

Hjelmås, E. and Low, B. K. 2001. Face detection: A survey. *Computer vision and imageunderstanding*, 83(3); 236–274.

Hochreiter, S. and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8); 1735–1780.

Howard, A., Zhmoginov, A., Chen, L.-C., Sandler, M., and Zhu, M. 2018. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Hu, J., Shen, L., and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*; 7132–7141.

Hussain, S. A. and Al Balushi, A. S. A. 2020. A real time face emotion classification and recognition using deep learning model. In *Journal of physics: Conference series*, volume 1432; 012087. IOP Publishing.

Ioffe, S. and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*; 448–456. PMLR.

Jorge-Martinez, D., Butt, S. A., Onyema, E. M., Chakraborty, C., Shaheen, Q., De-La- Hoz-Franco, E., and Ariza-Colpas, P. 2021. Artificial intelligence-based kubernetes container for scheduling nodes of energy composition. *International Journal of SystemAssurance Engineering and Management*; 1–9.

Kendon, A. 1983. The study of gesture: Some remarks on its history. In *Semiotics 1981*; 153–164. Springer.

Kingma, D. P. and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kollias, D., Schulc, A., Hajiyev, E., and Zafeiriou, S. 2020. Analyzing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*; 637–643. IEEE.

Kollias, D., Sharmanska, V., and Zafeiriou, S. 2019. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint*

arXiv:1910.11111.

Kollias, D., Sharmanska, V., and Zafeiriou, S. 2021. Distribution matching for heterogeneous multi-task learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*.

Kollias, D. and Zafeiriou, S. 2019. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*.

Kollias, D. and Zafeiriou, S. 2021. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*.

Kossaifi, J., Tzimiropoulos, G., Todorovic, S., and Pantic, M. 2017. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65; 23– 36.

LeCun, Y. 1998. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

Minaee, S., Abdolrashidi, A., Su, H., Bennamoun, M., and Zhang, D. 2019. Biometrics recognition using deep learning: A survey. *arXiv preprint arXiv:1912.00271*.

Minaee, S., Luo, P., Lin, Z., and Bowyer, K. 2021. Going deeper into face detection: Asurvey. *arXiv preprint arXiv:2103.14983*.

Nwankpa, C., Ijomah, W., Gachagan, A., and Marshall, S. 2018. Activation functions: Comparison of trends in practice and research for deep learning. *arXiv preprint arXiv:1811.03378*.

O'Shea, K. and Nash, R. 2015. An introduction to convolutional neural networks. *arXivpreprint arXiv:1511.08458*.

Pashine, S., Dixit, R., and Kushwah, R. 2021. Handwritten digit recognition using machine and deep learning algorithms. *arXiv preprint arXiv:2106.12614*.

Pease, B. and Pease, A. 2008. *The definitive book of body language: The hidden meaning behind people's gestures and expressions*. Bantam.

Ramzan, F., Khan, M. U. G., Rehmat, A., Iqbal, S., Saba, T., Rehman, A., and Mehmood, Z. 2020. A deep learning approach for automated diagnosis and multi- class classification of alzheimer's disease stages using resting-state fmri and residual neural networks. *Journal of medical systems*, 44(2); 1–16.

Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. 2013. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*; 1–8. IEEE.

Rosenstein, D. and Oster, H. 1988. Differential facial responses to four basic tastes in newborns. *Child development*; 1555–1568.

Simonyan, K. and Zisserman, A. 2014. Very deep convolutional networks for large-scaleimage recognition. *arXiv preprint arXiv:1409.1556*.

Suryanarayana, G., Chandran, K., Khalaf, O. I., Alotaibi, Y., Alsufyani, A., and

Alghamdi, S. A. 2021. Accurate magnetic resonance image super-resolution using deep networks and gaussian filtering in the stationary wavelet domain. *IEEE Access*, 9; 71406–71417.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. 2017. Inception-v4, inception- resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*; 1–9.

Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., and Le, Q. V. 2019. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2820–2828.

Udofia, U. 2018. Basic overview of convolutional neural network (cnn). *Retrieved May*,27:2019.

Wayman, J., Jain, A., Maltoni, D., and Maio, D. 2005. An introduction to biometric authentication systems. In *Biometric Systems;* 1–20. Springer.

Yang, S., Luo, P., Loy, C.-C., and Tang, X. 2016. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*; 5525–5533.

Zagoruyko, S. and Komodakis, N. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146*.

Zhang, Cha, and Zhengyou Zhang. "A survey of recent advances in face detection." (2010).

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10); 1499–1503.

Zhang, S., Pan, X., Cui, Y., Zhao, X., and Liu, L. 2019. Learning affective video features for facial expression recognition via hybrid deep learning. *IEEE Access*, 7; 32297– 32304.