REPUBLIC OF TÜRKİYE
HASAN KALYONCU UNIVERSITY
GRADUATE EDUCATION INSTITUTE
DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

ANALYSIS OF RACIAL BIAS IN FACIAL EMOTION RECOGNITION

Fatih ALİSİNANOĞLU

Ph. D. THESIS

IN

ELECTRICAL AND ELECTRONICS ENGINEERING

GAZİANTEP - 2024

GRADUATE EDUCATION INSTITUTE
DOCTORAL THESIS ACCEPTANCE AND APPROVAL FORM

Electrical and Electronics Engineering Department, student of the Doctorate of Electrical and Electronics Engineering program Fatih ALİSİNANOĞLU prepared and submitted the thesis titled "Analysis of Racial Bias in Facial Emotion Recognition" and defended successfully at the date of 11/06/2024 and accepted by the jury as a Ph.D. thesis.

Position	Title, Name and Surname	Department/University	Signature
Supervisor	Prof. Dr. M. Sadettin ÖZYAZICI	EEE / Hasan Kalyoncu Uni.	
Jury Member	Prof. Dr. Arif NACAROĞLU	EEE / Gaziantep Uni.	
Jury Member	Prof. Dr. M. Fatih HASOĞLU	AE / Hasan Kalyoncu Uni.	
Jury Member	Assoc. Prof. Dr. Bülent HAZNEDAR	COMPE / Gaziantep Uni.	
Jury Member	Dr. K. Sercan BAYRAM	EEE / Hasan Kalyoncu Uni.	

This thesis is accepted by the jury members selected by institute management board and approved by institute management board.

Doç. Dr. Ufuk AKBAŞ
Director

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Fatih ALİSİNANOĞLU

Date: 11/06/2024

HASAN KALYONCU ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
ELEKTRİK ELEKTRONİK MÜHENDİSLİĞİ BÖLÜMÜ

YÜZ İFADESİ TANIMADA IRKSAL ÖNYARGI ANALİZİ

Fatih ALİSİNANOĞLU

DOKTORA TEZİ

Danışman
Prof. Dr. M. Sadettin ÖZYAZICI

ÖZET

Çok boyutlu girdilere ve öznel etiketlere sahip derin öğrenme modellerinin karmaşıklığı, ortaya çıkan sonuçlarda tarafsızlık sorunlarını da beraberinde getirmektedir. Veri setlerinin ırksal olarak dengesiz dağılıma sahip olabildiği yüz ifadesi tanıma yapılarında, kullanılan modeller farklı ırk grupları için taraflı ve önyargılı sonuçlar üretebilmektedir. Bu durum, tarafsızlık konusundaki endişeleri artırmakta ve ırksal önyargı konusunda daha fazla araştırma yapılması ihtiyacını öne çıkartmaktadır. Bu tür, ırksal önyargıdan kaynaklanan hatalı genellemeler ve sınıflandırmalar, modellerin gerçek dünya performansını da olumsuz etkileyebilmektedir. Bu tezde, farklı ırk dağılımlarına sahip eğitim setleri oluşturmak amacı ile alt örnekleme teknikleri kullanılarak, ırksal önyargının değerlendirmesi ve ardından popüler yüz ifadesi tanıma metotlarının kullanımı ile ırksal önyargı konusunda test performanslarının incelenmesi amaçlanmıştır.

Anahtar Kelimeler: Yüz ifadesi tanıma, Derin sinir ağları, Irksal önyargı, Yapay zeka etiği.

HASAN KALYONCU UNIVERSITY
GRADUATE EDUCATION INSTITUTE
DEPARTMENT OF ELECTRICAL & ELECTRONICS ENGINEERING

ANALYSIS OF RACIAL BIAS IN FACIAL EMOTION RECOGNITION

Fatih ALİSİNANOĞLU

DOCTORAL THESIS

Supervisor
Prof. Dr. M. Sadettin ÖZYAZICI

ABSTRACT

The complexity of deep learning models with high-dimensional inputs and subjective labels extends to fairness. In facial emotion recognition, where datasets can be racially unbalanced, models may produce biased results for different racial groups. This raises concerns about fairness and highlights the need for further research on racial bias. Inaccurate generalization owing to such bias could negatively affect real-world performance. This thesis aims to evaluate racial bias using subsampling techniques to create training sets with diverse racial distributions and then examine test performances across these experiments using state-of-the-art facial emotion recognition techniques.

Keywords: Facial expression recognition (FER), Deep neural networks, Racial Bias, AI Ethics.

ACKNOWLEDGMENTS

I would like to acknowledge and give my warmest thanks to my supervisor Prof. Dr. M. Sadettin ÖZYAZICI who made this study possible. His guidance and advice carried me through all the stages of drafting my thesis. I would also like to thank my committee members for letting my defense be an enjoyable moment, and for your brilliant comments and suggestions, thanks to you.

I would also like to give special thanks to my family as a whole for their continuous support and understanding when undertaking my research and drafting my thesis. Your prayer for me was what sustained me this far.

Finally, I would like to thank God, for letting me through all the difficulties. I have experienced your guidance day by day. You are the one who let me finish my degree. I will keep on trusting you for my future.

Fatih ALİSİNANOĞLU

Gaziantep, 2024

Dedicated to my family…

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

FER	Facial Expression Recognition
ML	Machine Learning
AI	Artificial Intelligence
DL	Deep Learning
DNN	Deep Neural Network
RAI	Responsible Artificial Intelligence
XAI	Explainable Artificial Intelligence
CNN	Convolutional Neural Network
NLP	Natural Language Processing
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
MLP	Multilayer Perceptron
ReLU	Rectified Linear Units
GPU	Graphical Processor Unit
NN	Neural Network
SVM	Support Vector Machine
LDA	Linear Discriminant Analysis
MLR	Multinomial Logistic Regression
HMM	Hidden Markov Model
FACS	Facial Action Coding System
AU	Action Units
DCNN	Deep Convolutional Neural Network
BIF	Biologically Inspired Features
KCFA	Kernel Class-Dependent Feature Analysis
LBP	Local Binary Pattern
GAN	Generative Adversarial Network
SCN	Self-Cure Network
TL	Transfer Learning
ANN	Artificial Neural Network
CPU	Central Processor Unite
RAM	Random Accessible Memory
W	White

B	Black
A	Asian
I	Indian
IA	Indian & Asian
IB	Indian & Black
IW	Indian & White
AB	Asian & Black
AW	Asian & White
WB	White & Black
IAWB	Indian & Asian & White & Black

LIST OF SYMBOLS

L	Perceptron Layer
P	Prediction
ϕ	Activation function
$W^{(k)}$	Matrix
k	Layer
θ	Nonlinearity parameter
x	Input
y	Single dimension output
f	Application function
f_j	Artificial neuron function
b_j	Bias
α	Fixed parameter
$*$	Convolution operator
I	2D image
s	Stride
p	Image size
C_0	Kernel
W_i	Image width
H_i	Image height
C	Kernel
C^0	Number of kernels
t	Time
h_t	Output
$\widehat{Y}$	Predicted gender
Y	True gender
A	Demographic group
D	Race group
$T(\theta)$	Sampling grid
ψ	Output Layer
ℓ	Regularization
λ	Regularization weight

1. INTRODUCTION

1.1. General Overview

Emotion recognition, commonly known as facial expression recognition (FER), involves the identification and analysis of facial expressions displayed by individuals in images or videos [1]. This complex process comprises of three primary stages: face detection, feature extraction, and emotion classification.

Facial expressions are widely employed across various sectors, such as security, medicine, social sciences, marketing, and human-machine interaction, with the aim of achieving improved outcomes from numerous vantage points [2]. Consequently, Facial Expression Recognition has become a prominent area of research in computer vision and has garnered significant interest over the past two decades [2]. The use of deep learning techniques in computer vision tasks [3], particularly in facial expression recognition [4], has gained popularity owing to technological advancements [5] that have led to a more detailed analysis of annotated facial image datasets [6].

Historically, automatic emotion recognition relied on the extraction of domain-specific features such as facial action units. However, with the rapid progression of machine learning (ML) and deep learning (DL) techniques, deep neural networks (DNNs) have emerged as a prominent approach for developing facial emotion recognition models. Such models necessitate expansive datasets to ensure robustness and accuracy in their predictions. In recent years, researchers have proposed DL models which leverage more expansive datasets for training [6].

However, these techniques are vulnerable to biases in the data, which can result in inconsistent outcomes and reduced performance if the collected data does not adequately represent the diversity of real-life samples [7]. Moreover, biases towards particular races in facial expression recognition raise concerns about fairness, particularly in fields such as policing and surveillance, where criminal charges and prosecution are at stake [8]. Biases within DNNs can primarily be attributed to two fundamental sources: the training data and the algorithms themselves. Given that models learn from the input data, any biases present within the underlying datasets are inherently ingrained within the learning process of the algorithms [9]. Furthermore, the design of the feature extraction processes for these models may introduce biases that disproportionately affect different racial

groups [10]. An illustrative example is the consideration of skin color as a learned feature extracted during the deep-learning process, which can lead to unfair predictions [11].

1.2. Research Objectives

For a decade, the industry has released applications for facial detection [12]. These models claim to be more than 90% of accuracy; however, this does not seem to be the case when released to the public [13]. For example, Facebook, now META, released its facial detection model DeepFace in 2014, which was trained on over four million images and claimed to be more accurate at detecting faces than a human being, with an accuracy of 97.35%. This claim is misleading as it implies that the model's performance would be able to detect all types of people, but this was not the case [14]. The claim of human accuracy and 97.35% accuracy refers to the model's ability to detect faces more accurately than humans within the dataset used for training [15]. To further clarify this, humans may not have been able to see past occlusions in the dataset, or that the model's training set was racially biased. It was not publicized until 2020, when users noticed that Facebook's facial detection and labeling models auto-labeled Black men as primates [16].

The emergence of social concerns has led to the development of new academic terms, such as Responsible AI (RAI) and Explainable AI (XAI) [38]. Governments have acknowledged these concerns, resulting in laws and regulations that support them. The scientific community and businesses are actively working to address this significant issue [34]. The RAI framework focuses on promoting ethical AI use, whereas XAI emphasizes providing justification, explanations, and evidence for the decisions and behaviors of algorithms.

Ethical AI involves various aspects including the integration of robotics in society and digital privacy. One significant concept is algorithmic fairness, which refers to how systems can replicate human biases and discriminate against individuals based on protected characteristics such as sex, gender, race, or age [33]. While developing bias is a complex process, deep learning techniques are particularly susceptible to bias in datasets. These techniques learn patterns autonomously and often confuse patterns that correlate with the target class [38]. Consequently, models can incorporate and amplify correlations, leading to biased and differentiated predictions for certain individuals and races. To address these issues, it is crucial to measure the bias in both the final models

and datasets. Although measuring bias in source datasets has not received much attention, it is essential to validate new bias mitigation methods, explain bias transfer throughout the training process, and provide demographic descriptions of the environment in which a dataset or model can be safely used. One of the fundamental principles of ethical AI is fairness, which entails ensuring that AI algorithms do not display preferential treatment towards specific groups of people. This concept involves recognizing and addressing biases rather than dismissing them. Unfortunately, supervised machine learning algorithms, which rely entirely on training data, often inadvertently absorb hidden prejudices. This study aimed to investigate whether there is racial bias in deep learning techniques for FER. To determine whether a model exhibits bias, researchers typically need to conduct the same experiment on distinct test groups categorized by the target feature, which, in this case, is race. Furthermore, it is important to note that the results of these experiments can provide valuable insights into the potential sources of bias in the model and inform the development of more fair and equitable deep learning models in the case of FER.

The current literature on racial bias in FER reveals a lack of research in this area [7]. The datasets that are accessible to the public are inadequate in terms of racial diversity of the individuals included. Furthermore, the distribution of races within these datasets is uneven, and the methods used do not address the issue of racial bias or provide information on how to mitigate or prevent it. These factors present a significant challenge in the study of racial bias in FER [104]. To address this deficiency, a novel dataset comprising samples from a range of racial groups developed during this thesis study and a simulation study conducted to investigate the racial bias.

1.3. Thesis Outline

This thesis is comprised of seven chapters and each section provides a concise historical overview of the respective methodology. Chapter 2 explores the essential principles and the fundamentals of deep learning and facial expression analysis and also covers the facial expression recognition pipeline and past work, as well as a discussion of semi-supervised and self-supervised learning techniques, and their potential for expansion. Chapter 3 presents a comprehensive review of artificial intelligence ethics. The literature review in Chapter 4, concludes with an examination of the various biases present in a deep learning model. As the state-of-the-art deep learning architectures serves as the foundation of the

model used in this thesis, Chapter 5 explores this topic in greater detail. Chapter 5 also offers a review of the datasets employed in this study, both balanced and unbalanced, as well as information on the datasets used for training. Finally, Chapter 6 provides information on the training setup and the analysis of the results. This is followed by Chapter 7 in the conclusion and future work.

2. FUNDEMANTALS

Artificial intelligence, specifically deep neural networks, have revolutionized researchers' ability to create generalized solutions. This section discusses the fundamentals of deep learning to demystify the black box, which is a neural network. This section also reviews the fundamental learning techniques used in this study, namely transfer learning and semi-supervised self-supervised learning. The goal of this chapter is to provide a foundational background for reference through the remainder of this thesis, explore the history of these techniques, and explain the benefits and drawbacks of each approach.

The latter portion of this chapter is comprised of a comprehensive review of deep learning, commencing with the fundamentals of perception, and progressing to convolutional neural networks (CNN). Subsequently, the focus of this section will shift to the topic of transfer learning, a highly effective training technique that is widely utilized.

2.1. Review of Deep Learning

2.1.1. Description

Artificial intelligence (AI), which is exhibited by machines, has been proven to be an effective approach to human learning and reasoning. In 1950, the concept of AI was introduced through the proposal of "The Turing Test," which aimed to explain how a computer could perform human cognitive reasoning [28]. As a research field, AI has been further divided into more specific subfields. For example, Natural Language Processing (NLP) has been applied to enhance the writing experience in various applications. NLP is commonly divided into machine translation, which is the translation of languages, and other subfields. Machine translation algorithms have been developed to address both grammatical structures and spelling mistakes. Additionally, AI uses a set of words and vocabulary related to the main topic to suggest changes to the writer or editor. Figure 2.1 provides a detailed overview of how AI covers seven subfields of computer science.

Machine learning and data mining have garnered significant attention and are currently the most researched topics in the academic community [27]. These interdisciplinary fields focus on analyzing numerous possibilities for characterizing databases. The databases in question were collected for statistical purposes, and statistical curves can be used to

describe past and present behavior to predict future trends. Despite this, classic techniques and algorithms have been the only methods employed to process these data, whereas optimizing these algorithms could lead to effective self-learning. Implementing better decision-making processes based on existing values, multiple criteria, and advanced statistical methods can benefit various fields. One of the most significant applications of this optimization is in medicine, where symptoms, causes, and medical solutions generate large databases that can be used to predict better treatments [41].

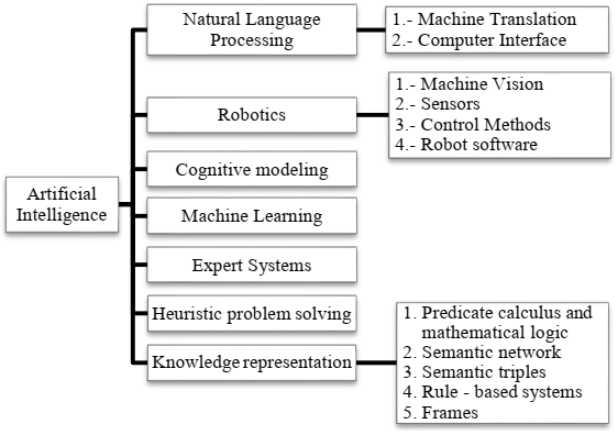


Figure 2.1: AI and related computer science subfields.

Numerous methodologies have been established owing to the extensive range of research encompassed by ML. These include clustering, Bayesian networks, deep learning, and decision-tree learning. This review focuses primarily on deep learning by exploring its core concepts and historical and current applications across various disciplines. In addition, it highlights the accompanying charts that depict the remarkable growth of deep learning research in recent years, as demonstrated by the significant increase in the number of publications in scientific databases.

Deep Learning (DL) emerged in 2006 as a novel research area in machine learning. Initially referred to as hierarchical learning, it typically encompasses numerous research domains associated with pattern recognition [7]. Deep learning primarily focuses on two critical factors: nonlinear processing in multiple layers or stages and supervised or unsupervised learning. Nonlinear processing in multiple layers pertains to an algorithm wherein the current layer utilizes the output of the preceding layer as the input. By establishing a hierarchy among the layers, the importance of the data is organized in a

manner deemed useful. Conversely, supervised, and unsupervised learning were linked to the class target label. The presence of a label denotes a supervised system, whereas its absence signifies an unsupervised system.

2.1.2. Applications

Deep learning involves the use of abstract layer analysis and hierarchical methods to improve results and optimize processing times in various computing tasks. This technique has various real-life applications, including digital image processing, where the coloring of grayscale images is performed manually by users based on their judgment [24]. However, with the use of deep-learning algorithms, coloring can now be performed automatically using a computer. Similarly, deep learning can be used to add sound to mute drumming videos using Recurrent Neural Networks (RNN) [23]. Deep learning has been applied in several fields including natural language processing, where it has been used for image caption generation and handwriting generation. Other applications of deep learning include digital image processing, medicine, and biometrics.

Deep learning is a powerful technique that enhances the outcomes and expedites processing times in numerous computing operations [20]. It finds extensive application in the domain of natural language processing, where it is utilized for tasks such as image caption generation and handwriting generation. In addition, deep learning is widely employed in digital image processing, medicine, and biometrics [30].

Natural language processing is a method that enables machines to comprehend natural language, which is a challenging task for humans, by analyzing its syntax, expressions, and semantics. This is achieved by training the machines to recognize linguistic nuances and generate suitable responses. In the legal field, the utilization of deep learning for document summarization has become commonplace, rendering paralegals unnecessary. Additionally, deep learning has been shown to be effective in several natural language processing tasks, such as question answering, language modeling, text classification, Twitter analysis, and sentiment analysis.

In the case of healthcare, the use of deep learning has experienced a surge in popularity owing to its integration with patient data and assessment of health conditions, including blood pressure, sugar levels, and heart rate. This technology is employed in medical

imaging solutions, chatbots that recognize patterns in patient symptoms, and deep learning algorithms that can discern specific types of cancer and rare diseases. The incorporation of deep learning in healthcare has endowed medical professionals with invaluable insights, empowering them to detect issues early and deliver more personalized care to patients.

Virtual assistants, including Alexa, Siri, and Google Assistant, are among the most prevalent applications of deep learning. These assistants employ deep learning to enhance their understanding of human speech and accents, thereby offering a more personalized experience. By analyzing natural language, they can learn about their users' preferences, ranging from their preferred restaurants to their favorite sports. In addition to converting speech to text and taking notes, virtual assistants can also arrange appointments and perform a variety of tasks, such as running errands or organizing chores for individuals and teams. These assistants can manage a broad range of tasks, including auto-responding to calls and organizing schedules.

Visual recognition is attempted to examine a collection of vintage photographs that evoke nostalgic memories. To begin, you must sort them manually, as there are no metadata available. As downloaded images often lack this information, the best option is to arrange them by date. Fortunately, advancements in Deep Learning now allow photographs to be categorized based on several factors such as location, facial recognition, groups of people, events, and dates. Innovative visual recognition techniques, including convolutional neural networks, TensorFlow, and Python, can be utilized to search for a specific photo within a vast library, such as Google’s extensive picture collection. These advanced methods have shown promising results in the field of visual recognition, particularly for digital media management.

The creation of sounds for silent videos involves the use of convolutional neural networks and long short-term memory (LSTM) recurrent neural networks [4]. A deep learning model typically matches video frames with a database of prerecorded sounds to select the most appropriate sounds for a specific scene. This process is conducted by training 1000 videos that depict drumsticks striking different surfaces and producing a variety of sounds. Deep learning models then utilize these videos to determine the best

sound for a given video. To achieve the best results for predicting whether a sound is real or artificial, a turning test-like setup was established.

Sentiment analysis is a method that employs natural language processing, text analysis, and statistical techniques to evaluate and analyze client sentiment. By collecting customer feedback from various sources, including social media platforms like Twitter and Facebook, as well as structured data such as surveys, consumer feedback, and call center data, sentiment analysis can be performed. Unstructured data, on the other hand, refers to information that is not owned by a corporation or individual, such as data collected by independent sources. Deep learning is a critical component of sentiment analysis as it enables the categorization of sentiments and the extraction of opinions and emotions from text.

The colorization of grayscale images is another function of deep-learning networks. The task of adding color to these pictures is known as image colorization, which has historically been performed manually by skilled professionals. However, with the aid of deep learning, images can now be colored by utilizing objects and their surrounding context within the photograph, like how a human would. This is a highly impressive visual feature that leverages advanced and extensive convolutional neural networks developed for image colorization, such as ImageNet. Typically, this process requires the application of large convolutional neural networks and supervised layers to generate a colored image.

Automatic language translation is a process that entails converting words, phrases, or sentences from one language to another. Although this technology has existed for quite some time, deep learning has proven to be the most effective method in two critical areas: automatic text translation and image translation. Deep learning enables text translation without preprocessing, allowing the algorithm to learn the relationships between words and how they are transferred to a new language. This was accomplished using stacked networks of large LSTM recurrent neural networks, which were then translated. Convolutional neural networks are employed to recognize letters in images and their locations within a scene. Once these elements have been identified, they can be converted into text, translated, and reproduced as an image with the translated text, a process commonly referred to as "immediate visual translation."

Deep learning, a subfield of machine learning that shows significant potential for future growth, has garnered increasing attention in recent years because of its unparalleled efficiency in enhancing machine functionality [28]. This application of artificial intelligence demonstrates remarkable capabilities, surpassing human intelligence, which makes it a critical component in shaping a better future. It is imperative to invest in and promote deep learning to maximize its benefits to the greater good of humanity.

2.1.3. The Perceptron Algorithm

The Perceptron was initially introduced by Frank Rosenblatt, a researcher at the Cornell Aeronautics Laboratory in 1957. Drawing from the earliest concepts of artificial neurons, he developed the "Perceptron learning rule," which serves as the foundation for Perceptron operation [42]. Perceptron is a neural network unit that function as artificial neurons and are responsible for processing and detecting patterns in input data. They are used to classify binary data and are trained using a supervised learning algorithm. The Perceptron learning rule allows artificial neurons to process and learn features in a dataset, making it a vital tool in machine-learning projects. This rule is commonly employed to classify data and supervise the learning capabilities of the binary classifiers. It is essential to mention that supervised learning involves teaching an algorithm to make predictions by providing it with data that has already been correctly labeled.

To grasp the theory of artificial neural networks, it is important to initially comprehend the role of perceptron. The human brain is comprised of a vast network of interconnected neurons that facilitate the processing and transmission of electrical and chemical signals. In contrast, artificial neurons are mathematical functions that mimic the structures of biological neurons [30]. They received data, processed it, calculated the sum, and produced a result using a nonlinear function. An artificial neural network is composed of multiple layers, including the input, output, and hidden layers, with data flowing from one layer to another. To function effectively, each neuron must receive data, which are processed and transmitted to the next neuron via synapses or connections between neurons. Dendrites, which are branches of neurons, receive information from other neurons, while the cell nuclei process the information received by the dendrites. Consequently, the perceptron is a critical component of deep learning neural networks.

The Perceptron Learning Rule dictates that an algorithm automatically determines the optimal weight coefficients for learning. By multiplying the characteristics of the input data by these weights, the algorithm can determine whether a neuron will "light up" or not. The Perceptron learning approach as shown in Figure 2.2, processes multiple input signals, and if the total number of these signals surpasses a specific threshold, an output signal is produced. Conversely, if the total number of signals does not meet the threshold, no output is generated. This process allows the prediction of a data sample category using the supervised learning approach of Machine Learning.

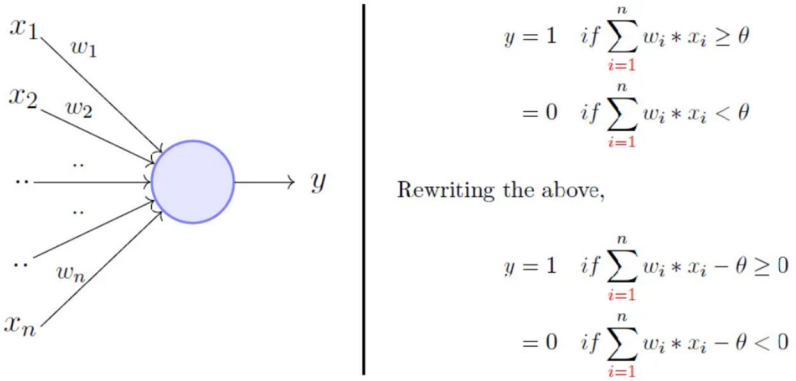


Figure 2.2: The perceptron learning model.

The Perceptron Learning Rule specifies that the algorithms automatically determine the ideal weight coefficients for learning. By multiplying the features of the input data with these weights, the algorithm can determine whether a neuron will be activated. The Perceptron processes multiple input signals: if the total number of these signals exceeds a specific threshold, an output signal is generated. Conversely, if the total number of signals does not meet the threshold, no output is generated. This process allows the prediction of a data sample category using a supervised learning approach in Machine Learning [18].

The perceptron is a widely used computational model (see Figure 2.3) that has been designed to process inputs, perform a weighted sum, and return a binary output of 1 if the sum is greater than a predetermined threshold and 0 otherwise. This general model is often employed in machine learning and artificial intelligence applications, and its operation is based on straightforward mathematical algorithms [19]. To accommodate

various application requirements, the threshold and weighting factors can be adjusted to satisfy specific requirements.

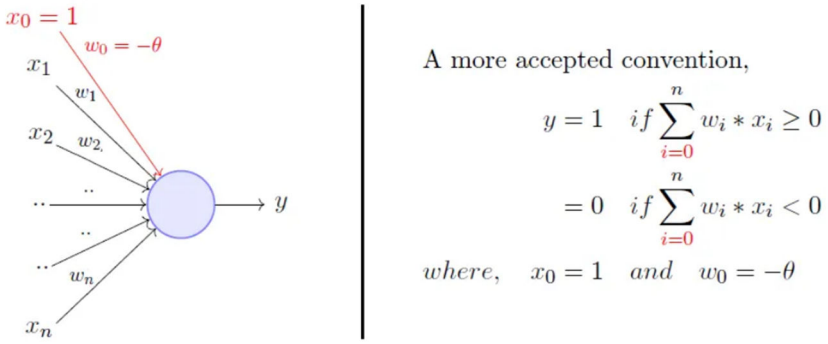


Figure 2.3: Computational model of the perceptron.

A single perceptron is limited to implementing linearly separable functions. It processes both real and Boolean inputs, assigning a set of weights and a bias to them.

A multilayer perceptron (MLP), (see Figure 2.4) also known as a neural network, is a complex structure comprising multiple concealed layers of neurons, with the output of one neuron serving as the input for the next layer. This process can also occur within the same layer or with neurons from previous layers (as seen in recurrent neural networks). The final layer, known as the output layer, employs distinct activation functions depending on the problem being addressed (regression or classification). Figure 2.4 below illustrates a neural network with three input variables, one output variable, and two hidden layers.

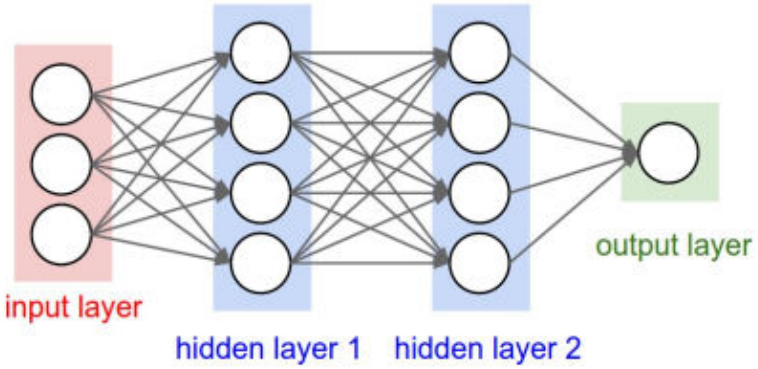


Figure 2.4: Multilayer perceptron model.

Multilayer perceptrons possess a unique structure in which each neuron in a layer is interconnected with all the neurons in the following layer but has no connection with those in the same layer [23]. The user is responsible for selecting the number of hidden layers and neurons in each layer, which are considered as parameters of the architecture. In addition, the activation functions for each layer can be chosen by the user. In the case of the output layer, the activation function is typically different from that used for hidden layers. In regression, no activation function is typically applied to the output layer. For binary classification, the output provides a prediction of $P(Y=1|X)$ and the sigmoid activation function is often used. For multiclass classification, the output layer has one neuron per class, which provides a prediction of $P(Y=i|X)$ for each class. The sum of these values must equal to 1, and the multidimensional softmax function is typically used for this purpose.

$$softmax(z)_i = \frac{exp(z_i)}{\sum_j exp(z_j)}.$$

The following description provides a concise summary of the mathematical formulation for a multilayer perceptron with L hidden layers:

We set $h^{(0)}(x) = x$.

For $k = 1, \ldots, L$ (*hidden layers*) ,

$$\alpha^{(k)}(x) = b^{(k)} + W^{(k)} h^{(k-1)}(x)$$
$$h^{(k)}(x) = \phi(\alpha^{(k)}(x))$$

For $k = L + 1$ (*output layer*) ,

$$\alpha^{(L+1)}(x) = b^{(L+1)} + W^{(L+1)} h^{(L)}(x)$$
$$h^{(L+1)}(x) = \psi(\alpha^{(L+1)}(x)) := f(x, \theta) .$$

the expression ϕ denotes the activation function, whereas ψ the output layer activation function, is designated for multiclass classification. At every stage, $W^{(k)}$ represents a matrix with a specified number of rows, neurons pertaining to layer k, and columns comprising neurons of layer $k-1$.

The mathematical formulation of a multilayer perceptron with L hidden layers is a complex process that involves numerous calculations and mathematical equations. The mathematical formulation of a multilayer perceptron is based on a series of linear equations that are used to train the neural network.

The first step in the mathematical formulation of a multilayer perceptron is to define the input layer, which consists of a set of input variables. These input variables were then multiplied by a set of weighted coefficients, which were used to calculate the output of the neural network. The output of the neural network was then passed through a series of hidden layers, each of which consisted of a set of neurons that were activated by a set of input variables.

The activation function used in the mathematical formulation of a multilayer perceptron is typically a sigmoid function that maps the output of the neural network to a range between 0 and 1. This function was used to ensure that the output of the neural network was bounded and did not exceed a certain value.

The final output of the mathematical formulation of a multilayer perceptron is the result of the activation function applied to the output of the last hidden layer. This output is then compared with the desired output to determine the error between the two. The error is then used to adjust the weighted coefficients of the neural network, which are updated iteratively until the error between the predicted output and desired output is minimized.

Overall, the mathematical formulation of a multilayer perceptron with L hidden layers is a complex process that requires a deep understanding of the linear equations, activation functions, and weighted coefficients. This is a fundamental process in the field of artificial intelligence and is used to train neural networks for a wide range of applications [24].

2.1.4. Hyperparameters

A machine learning model is a mathematical construct that comprises various adjustable parameters. These parameters must be determined by analyzing the data, and they can be adjusted to fit the available information through the training process. However, there is another type of parameter, called hyperparameters, that cannot be directly learned from standard training. These parameters are typically established before the training process

begins, and represent important aspects of the model, such as its complexity or the rate at which it should learn. The objective of this study was to investigate different strategies for optimizing hyperparameters in machine learning models.

The process of optimizing hyperparameters is essential for selecting the most effective settings for a machine learning model's hyperparameters. These settings, such as the learning rate, number of neurons in a neural network, and kernel size in a support vector machine, influence the learning process of the model. The objective of hyperparameter optimization is to identify the values that result in the best performance for a specific task. In the realm of machine learning, hyperparameters are configuration variables that are established prior to the commencement of the training process and that control the learning process rather than learning from the data. They play a critical role in optimizing the performance of a model and can significantly impact its accuracy, generalization, and other performance metrics.

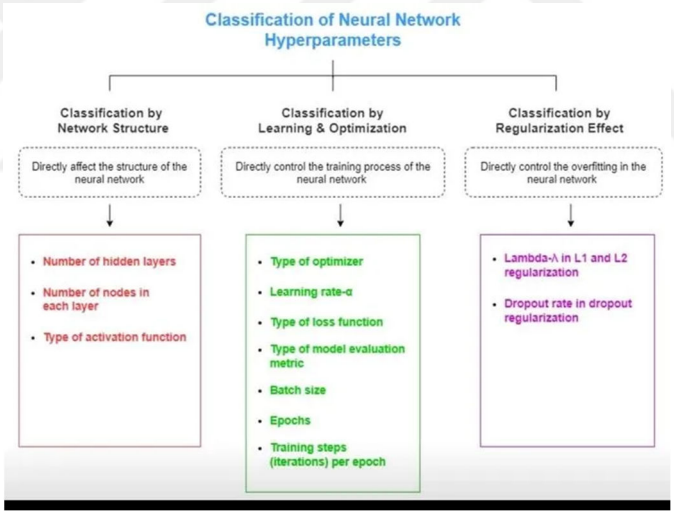


Figure 2.5: Classification of neural network hyperparameters.

Hyperparameters (as shown in Figure 2.5), serve a crucial function in governing the learning process of machine-learning models. These differ from the model parameters, which are the weights and biases learned from the data. There are several types of hyperparameters, including those used in neural networks. The essential hyperparameters for neural networks include the following.

1. Learning rate: This hyperparameter regulates the step size of the optimizer during each training iteration. An improper learning rate can lead to a slow convergence, instability, and divergence.

2. Epochs: This hyperparameter signifies the number of times the entire training dataset is passed through the model during the training. Increasing the number of epochs can enhance the model's performance, but also increase the risk of overfitting if not manage with care.

3. Number of layers: This hyperparameter determines the depth of the model and significantly affects its complexity and learning ability.

4. Number of nodes per layer: This hyperparameter determines the width of the model and influences its capacity to represent intricate relationships in data.

5. Architecture: This hyperparameter determines the overall structure of the neural network, including the number of layers, neurons per layer, and connections between layers. The optimal architecture depends on the complexity of the task and the size of the dataset.

6. Activation function: This hyperparameter introduces nonlinearity into the model, allowing it to learn the complex decision boundaries. Common activation functions include sigmoid, tanh, and rectified linear units (ReLU).

2.1.5. Deep Neural Networks

Deep learning is a collection of machine learning techniques that rely on intricate neural network configurations to process complex data and incorporate various nonlinear transformations [4]. These methods have witnessed considerable progress in areas such as audio and image processing, facial recognition, speech recognition, computer vision, and automated language processing, including text classification (e.g., spam detection). The potential applications of this technology are vast, as evidenced by the AlphaGo program, which uses deep learning to defeat the world champion in Go in 2016.

There are several neural network architectures including multilayer perceptrons, convolutional neural networks, and recurrent neural networks [23]. Multilayer

perceptrons are the oldest and simplest types of neural networks, whereas convolutional neural networks are specifically designed for image processing, and recurrent neural networks are used for sequential data such as text or time series. These architectures are constructed on a deep hierarchy of layers and require multiple optimization algorithms, initialization techniques, and careful selection of structures to achieve optimal results. Despite their impressive outcomes, deep learning methods currently lack a solid theoretical foundation [32].

They exhibit a sophisticated deep structure comprising numerous layers, which necessitates the utilization of advanced stochastic optimization algorithms in addition to strategic initialization and structure selection. Despite the absence of extensive theoretical foundations, they have consistently displayed exceptional performance [31].

An artificial neural network is a tool that exhibits nonlinear behavior with respect to its parameter θ, which is linked to an input x and the corresponding output $y=(x;\theta)$. For simplicity, we assume that y is a single-dimensional output, although it can also be multidimensional. The application of f adheres to a precise form. Neural networks are used for regression and classification purposes. As is common in statistical learning, parameter θ is determined using a learning sample. The function to minimize is not always convex, leading to the existence of local minimizers. The widespread acceptance of this approach is because of the universal approximation theorem. Le Cun (1986) introduced an efficient method for computing the gradient of a neural network, known as backpropagation of the gradient, which allows for easy acquisition of a local minimizer of the quadratic criterion.

An artificial neuron is a function of the f_j input $x=\left(x_1,\ldots,x_d\right)$ weighted by a vector of connection weights $w_j=\left(w_{j,1},\ldots,w_{j,d}\right)$, completed by a neuron bias b_j, and associated to an activation function ϕ, namely.

$$y_j=f_j(x)=\phi\left(\langle w_j,x\rangle+b_j\right)$$

Several activation functions can be considered.

- The identity function $\phi(x)=x$.

- The sigmoid function (or logistic) $\phi(x) = \frac{1}{1 + exp(-x)}$.
- The hyperbolic tangent function ("tanh") $\phi(x) = \frac{exp(x) - exp(-x)}{exp(x) + exp(-x)} = \frac{exp(2x) - 1}{exp(2x) + 1}$

Figure 2.6 is schematic representation of an artificial neuron where $\sum = \langle w_j, x \rangle + b_j$.

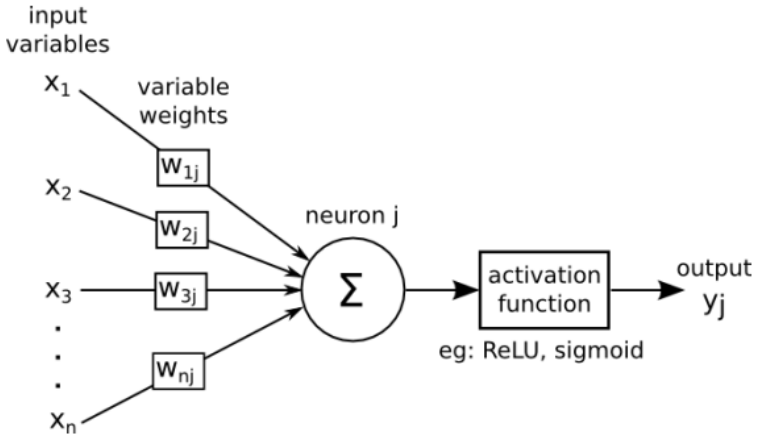


Figure 2.6: Schematic representation of an artificial neuron.

Figure 2.7 represents the activation functions desired above.

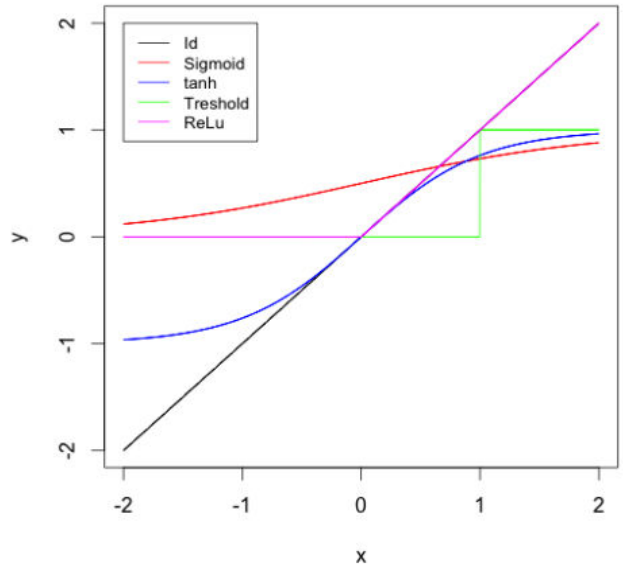


Figure 2.7: Activation functions.

Historically, the sigmoid function has been widely utilized because of its differentiable nature and ability to maintain values within the interval. Despite these advantages, the

sigmoid function presents a challenge in optimization because the gradient approaches zero when $|x|$ is not close to 0 making it difficult to optimize. The sigmoid function and its derivatives are shown in Figure 2.8 below.

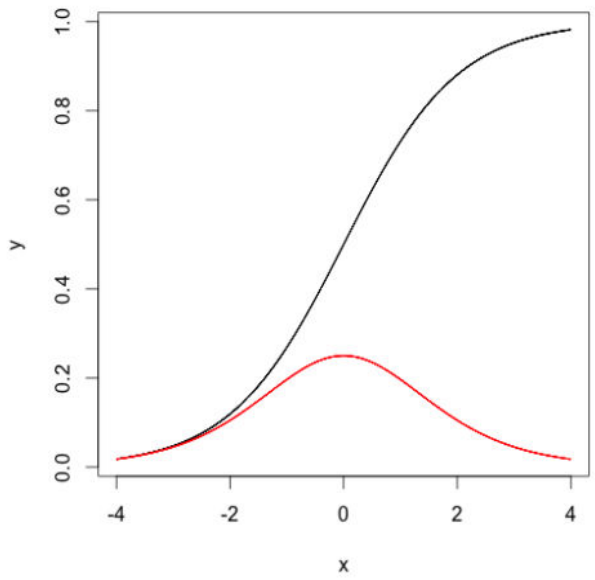


Figure 2.8: Derivatives (red) of the sigmoid function (black).

Deep learning utilizes neural networks with numerous layers, which presents challenges for the backpropagation algorithm to accurately estimate parameters. This challenge has led to the replacement of the sigmoid function with a rectified linear unit (ReLU) function. Although the ReLU function is not differentiable at 0, this practical limitation is not problematic because the probability of obtaining an entry equal to 0 is typically negligible. Furthermore, the ReLU function exhibits a sparsity effect, implying that the derivative is equal to 0 for negative values, resulting in no information being obtainable for such units. To address this issue, it is recommended that a small positive bias be added to ensure that each unit remains active. Additionally, various modifications to the ReLU function were considered to ensure that all units possessed a nonvanishing gradient, and the derivative was not equal to 0 for $x<0$. Namely

$$\varnothing(x) = max(x, 0) + \alpha\, min(x, 0)$$

Where α is either a fixed parameter set to a small positive value, or a parameter to estimate.

2.1.6. Convolutional Neural Networks

Some types of data, particularly images, are not well-suited for multilayer perceptrons. This is because of the models which are designed to process vector data, and images must be transformed into vectors to be used with these models, which can result in the loss of spatial information such as forms. Prior to the development of deep learning for computer vision, image processing relied on the manual extraction of variables of interest, known as features. However, these methods require significant expertise for image processing. The introduction of convolutional neural networks by LeCun [29] has revolutionized image processing and eliminated the need for manual feature extraction. CNNs operate directly on matrices or even tensors for images with three RGB color channels and are widely used for image classification (see Figure 2.9), image segmentation (see Figure 2.10), object recognition, and face recognition [56].

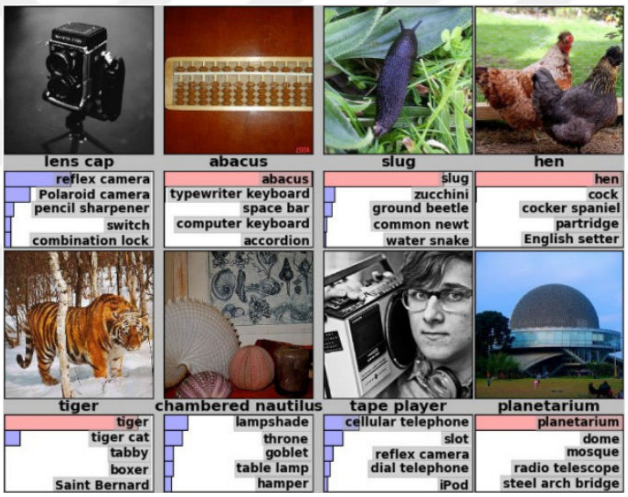


Figure 2.9: Image annotation.

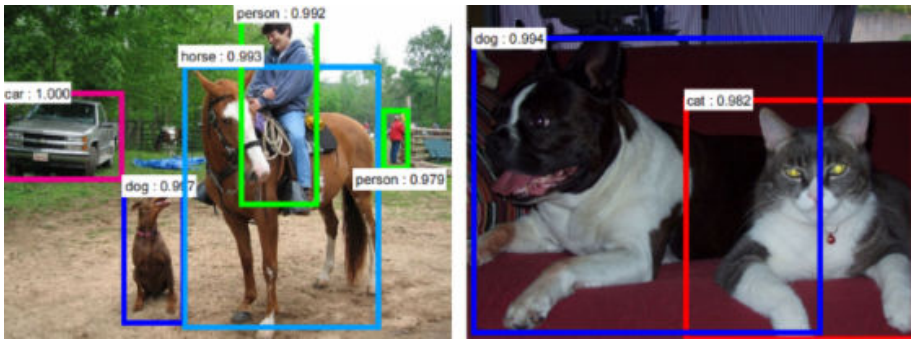


Figure 2.10: Image Segmentation.

A. Layers in a CNN

The discrete convolution between the two functions f and g is defined as

$$(f * g)(x) = \sum_t f(t)\, g(x+t)$$

2-dimensional signals such as images, 2D-convolutions are considered

$$(K * I)(i, j) = \sum_{m,n} K(m, n)\, I(i+n, j+m)$$

K is convolution kernel applied to a 2D signal (or image) I.

The process of two-dimensional convolution involves moving a convolution kernel across an image, as illustrated in the Figure 2.11 below. At every position, the kernel was convolved with the portion of the image being processed, resulting in a convolution value. The kernel is then shifted by a specific number of s pixels, s is known as stride. When the stride is small, redundant information is obtained. Additionally, zero padding, which entails adding a border of zero values around an image of size p, is used to control the output size. Suppose C_0 kernels (also called filters), each measuring $k \times k$ in size, are applied to an image with dimensions $W_i \times H_i \times C_i$ (where W_i denotes the width, H_i the height, and C_i the number of channels, typically $C_i = 3$). In this case, the output volume is $W_0 \times H_0 \times C_0$, where C_0 corresponds to the number of kernels considered.

$$W_0 = \frac{W_i - k + 2p}{s} + 1$$

$$H_0 = \frac{H_i - k + 2p}{s} + 1$$

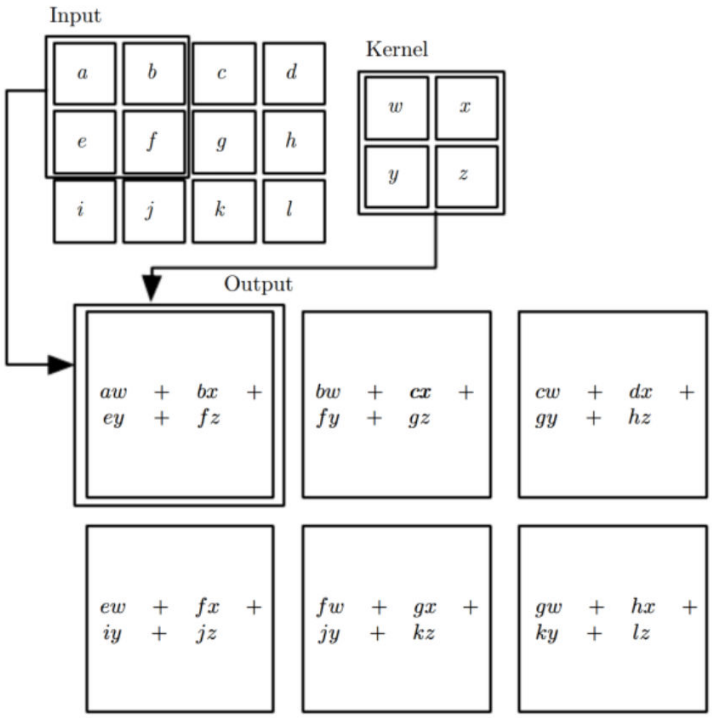


Figure 2.11: 2D convolution.

If an image contains 3 channels, the convolution with kernel $K_1\left(l=1,\ldots,C_0\right)$, which is a $5\times 5\times 3$ kernel where 3 corresponds to the number of channels in the input image, can be expressed using the following formula when applied to image I.

$$K_l * I(i,j) = \sum_{c=0}^{2}\sum_{n=0}^{4}\sum_{m=0}^{4} K_l(n,m,c)\, I(i+n-2, i+m-2, c)$$

Dealing with images containing C^i channels typically require a kernel in the form $\left(k,k,C^i,C^0\right)$, where C^0 signifies the number of output channels or kernels being considered. As shown in figure below, the specified values are $(5,5,3,2)$. The number of parameters connected to the kernel of shape $\left(k,k,C^i,C^0\right)$ is calculated as $(k\times k\times C^i+1)\times C^0$. These convolution operations are often accompanied by an activation function ϕ, which is typically a ReLU activation function. When using a kernel K of size $k\times k$ and an input image patch x of size $k\times k$, the resulting activation is obtained by sliding the $k\times k$ window and computing $(k\times k\times C^i+1)\times C^0$, where b represents a bias.

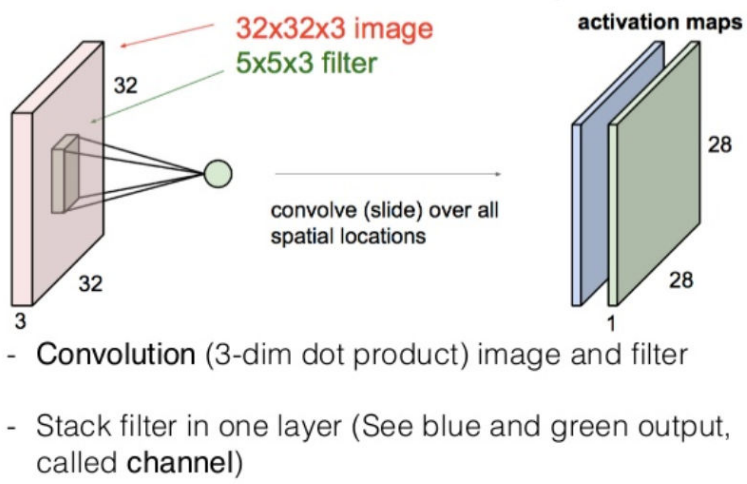


Figure 2.12: Convolution layer.

The convolution layer (as shown in Figure 2.12) is where the strength of a CNN lies, as it enables the network to learn the most useful filters (or kernels) for the specific task at hand, such as classification. Furthermore, multiple convolution layers can be employed, with the output of one convolution layer serving as the input for the next layer. This process allows the extraction of increasingly complex features as the network progresses.

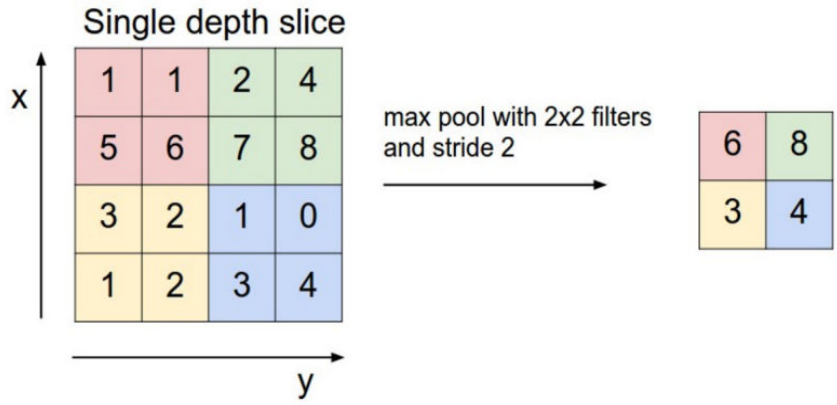


Figure 2.13: Effect of maxpooling on dimension.

Convolutional Neural Networks also consist of pooling layers, which lower the image dimensionality by calculating the average or maximum value of patches within the image. These layers operate on small parts of the image and include a stride like that of the convolutional layers. For instance, as shown in Figure 2.13, when using 2×2 patches and a stride of two, the output layer is determined by taking the maximum value. By dividing the width and height of the image by two, the dimensions of the image were reduced.

Although convolutional layers can also decrease the dimension, pooling layers offer the advantage of making the network less sensitive to minor changes in the input images. After implementing several convolution and pooling layers, a Convolutional Neural Network typically comprises a series of fully connected layers. The resulting tensor was then transformed into a vector augmented with additional perceptron layers.

B. Architectures of CNN

Several types of layers comprise a Convolutional Neural Network. The selection of an architecture is an overly complex process that involves more engineering than scientific precision. Consequently, it is essential to study architectures that have proven effective in the past and draw inspiration from well-known examples. In a traditional CNN, a series of convolutional layers are followed by pooling layers with fully connected layers added at the end. The LeNet network, as shown in Figure 2.14, is proposed by the inventor of the CNN, Yann LeCun [30], is an example of this type of architecture. Originally designed for digit recognition owing to computer limitations at the time, this network is composed of only a few layers and filters. However, since then, the network has been modified and expanded to include more layers and filters, allowing it to perform a wider range of tasks such as image classification and object detection.

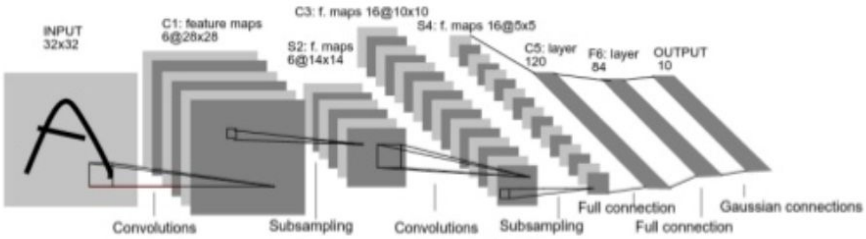


Figure 2.14: Architecture of LeNet.

With the advent of GPU (Graphical Processor Unit) cards, more intricate architectures have been proposed for CNNs such as AlexNet. This network won the ImageNet competition and is illustrated in a simplified version in Figure 2.15 below [23]. This competition involved classifying one million color images into 1000 distinct categories, with each image having a resolution of 224×224 pixels. AlexNet is comprised of 5 convolutional layers, 3 max-pooling 2×2 layers, and fully connected layers. As shown in

Figure 2.16, the first convolutional layer's kernel shape is $(11, 11, 3, 96)$ with a stride of $s = 4$ and the first output shape is $(55, 55, 96)$.

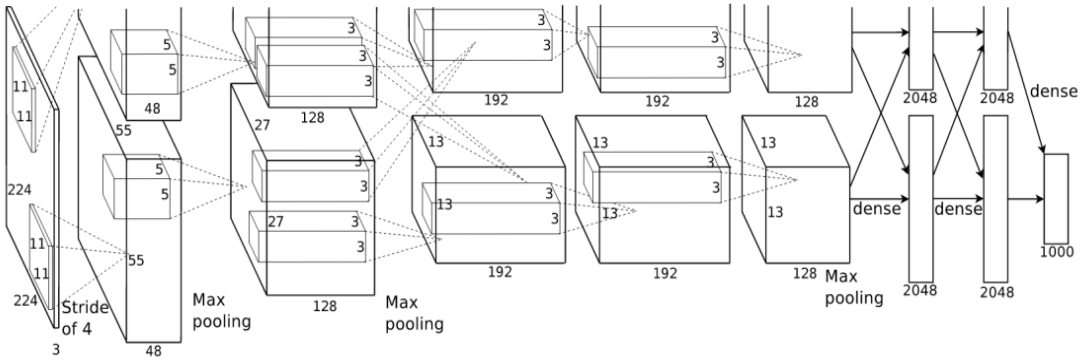


Figure 2.15: AlexNet architecture. Krizhevsky, A. et al (2012).

Input	227 * 227 * 3			
Conv 1	55*55*96	96	11 *11	filters at stride 4, pad 0
Max Pool 1	27*27*96		3 *3	filters at stride 2
Conv 2	27*27*256	256	5*5	filters at stride 1, pad 2
Max Pool 2	13*13*256		3 *3	filters at stride 2
Conv 3	13*13*384	384	3*3	filters at stride 1, pad 1
Conv 4	13*13*384	384	3*3	filters at stride 1, pad 1
Conv 5	13*13*256	256	3*3	filters at stride 1, pad 1
Max Pool 3	6*6*256		3 *3	filters at stride 2
FC1	4096	4096	neurons	
FC2	4096	4096	neurons	
FC3	1000	1000	neurons	(softmax logits)

Figure 2.16: Tabular representation of AlexNet network architecture.

Figure 2.17 shows another example of deep convolutional networks for large-scale image recognition.

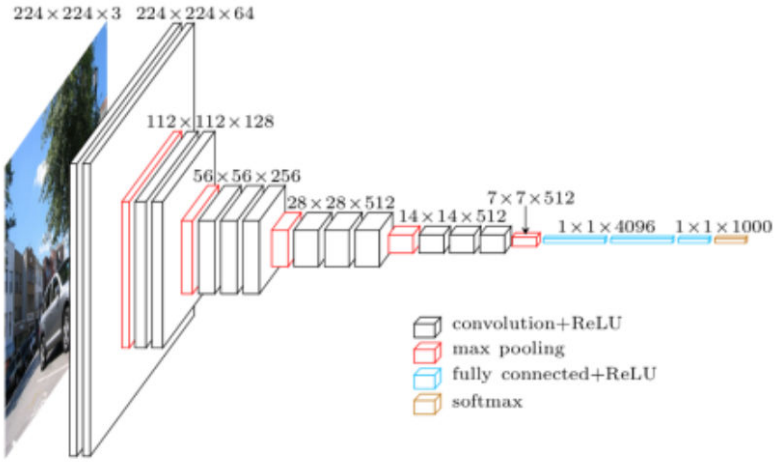


Figure 2.17: Deep convolutional networks example

The revolutionary GoogleNet network [23], which emerged victorious in the 2014 competition, was a trailblazing type of convolutional neural network (CNN) that incorporated not only convolution and pooling layers, but also groundbreaking modules called Inception as depicted in Figure 2.18.

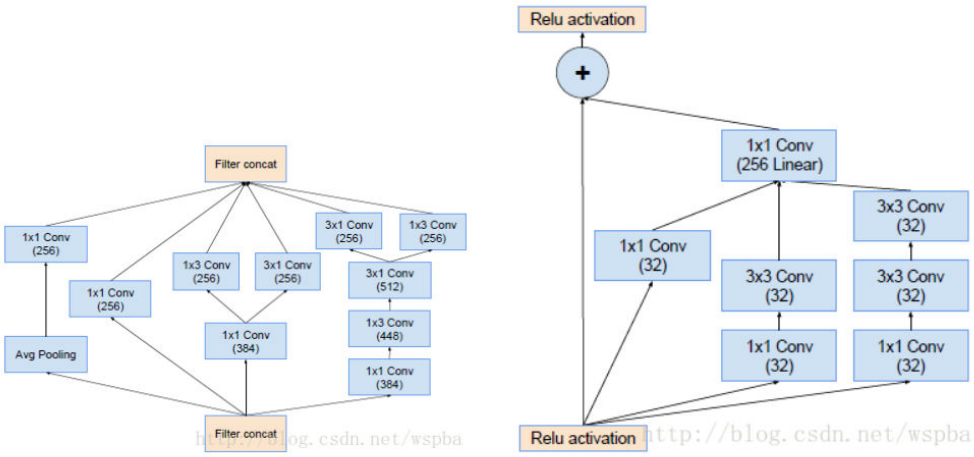


Figure 2.18: Modules of Inception network [81].

The novel aspect of ResNet is the addition of a connection between the input and output of a layer (or group of layers) [81] to reduce the number of parameters as shown in Figure 19. Unlike conventional CNNs, ResNet does not have fully connected layers. Although GoogleNet and ResNet are significantly deeper than the previous CNNs, they possess fewer parameters. However, they require more memory than traditional CNNs, such as VGG or AlexNet.

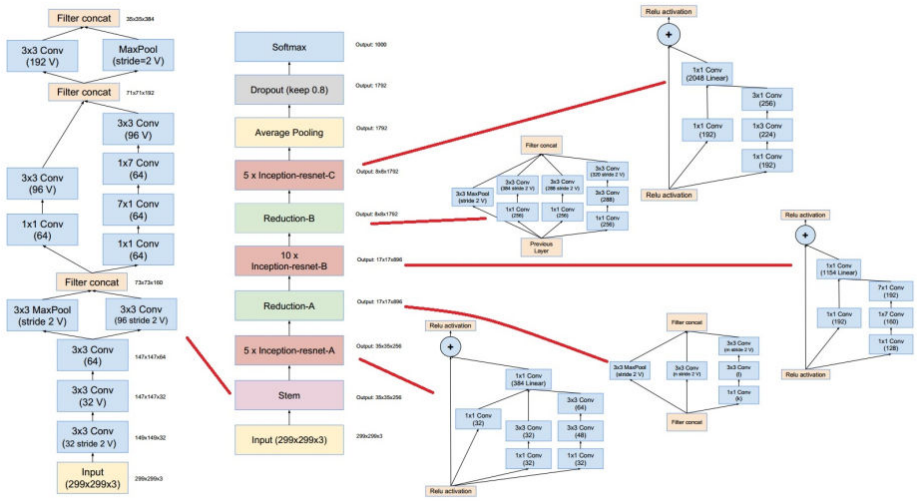


Figure 2.19: Inception-v4, Inception-resnet [81].

Figure 2.20 presents a side-by-side comparison of the depth and performance of the various networks in the ImageNet challenge.

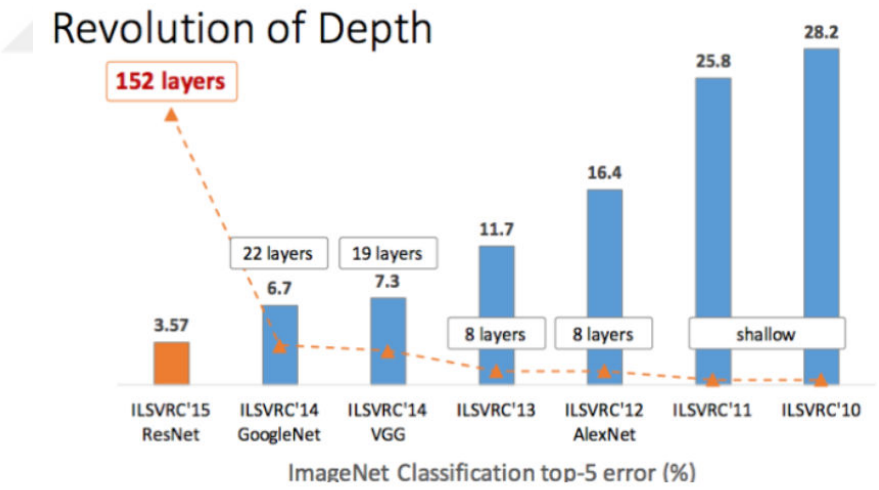


Figure 2.20: Comparison of the various networks in the ImageNet challenge.

2.1.7. Recurrent Neural Networks

Simple Recurrent Networks were initially proposed by Elman (1990) [21] and Jordan (1990) [24]. Elman's recurrent network is a multi-layer perceptron with a single unit layer that loops back. We denote the input at time t as $x(t)$, output at time t as $\hat{y}(t)$, and hidden layer at time t as $\hat{z}(t)$. For the kth component of the output:

$$\hat{y}^{(k)}(t) = \sum_{i=1}^{I} W_i^{(k)} \hat{z}_i(t) + b^{(k)}$$

$$\widehat{z}_i(t) = \sigma\left(\sum_{j=1}^{J} w_{i,j} x_j(t) + \sum_{l=1}^{I} \tilde{w}_{i,l} \widehat{z}_1(t-1) + b_i\right)$$

the activation function σ is applied to the neurons of the hidden layer that are connected to themselves, which are referred to as the context units. In Jordan's model, the last equation replaces $\widehat{z}_1(t-1)$ with $\hat{y}_1(t-1)$ and the context units are the output neurons. These models were originally developed for linguistic analysis and have since become widely used in natural language processing. Although the basic version of recurrent neural networks can capture long-term dependencies, new architects have been developed to address this issue.

Recurrent Neural Networks are traditionally employed to process sequential data such as text or time series. The earliest versions of RNNs were developed in the 1980s, featuring a hidden layer at time t which is determined by the input at time t, x_t and the previous hidden layer at time $t-1$ or the output at time $t-1$. This creates a loop between the hidden layer and itself, or between the output and hidden layers, as shown in Figure 2.21.

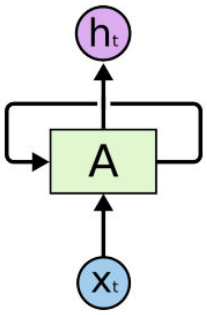


Figure 2.21: Diagram of an RNN.

While RNN may seem different from conventional neural networks at first glance, they consist of multiple copies of the same network, as demonstrated in Figures 2.21 and 2.22, each passing information to a subsequent unit. Figure 2.22 depicts the unrolled representation of the RNN.

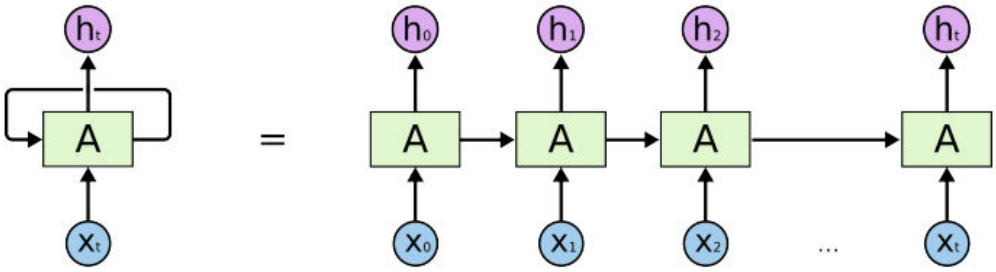


Figure 2.22: Representation of an RNN.

2.1.8. Long Short-Term Memory

In recent years, Recurrent Neural Networks have achieved success in various applications, such as speech recognition, translation, and image captioning. This success can primarily be attributed to the effectiveness of Long Short-Term Memory (LSTM), which is a specific type of RNN. LSTM cells, first introduced by Hochreiter and Schmidhuber (1997) [27], are designed to learn long-term dependencies. An LSTM cell comprises the state C_t and output h_t at time t. The inputs to the cell at time t include x_t C_{t-1}, and h_{t-1}. Within an LSTM, information is transmitted through gates that define the computations. These computations are governed by the following equations [27]:

$u_t = \sigma\left(W^u h_{t-1} + I^u x_t + b^u\right)$	Update gate H
$f_t = \sigma\left(W^f h_{t-1} + I^f x_t + b^f\right)$	Forget gate H
$\tilde{C}_t = tanh\left(W^c h_{t-1} + I^c x_t + b^c\right)$	Cell candidate H
$C_t = f_t \odot C_{t-1} + u_t \odot \tilde{C}_t$	Cell output H
$o_t = \sigma\left(W^o h_{t-1} + I^o x_t + b^o\right)$	Output gate H
$h_t = o_t \odot tanh\left(C_t\right)$	Hidden output H
$y_t = sof\,tmax\left(Wh_t + b\right)$	Output K
$W^u,\ W^f,\ W^c,\ W^o$	Recurrent weights $H \times H$
$I^u,\ I^f,\ I^c,\ I^o$	Input weights $N \times H$
$b^u,\ b^f,\ b^c,\ b^o$	Biases H

The differences between a classical RNN and LSTM can be clearly seen in Figure 2.23. While a standard RNN has a simple module A with only one layer, the repeated module in an LSTM consists of four layers (shown in yellow) that interact with each other, as described by the equations mentioned earlier.

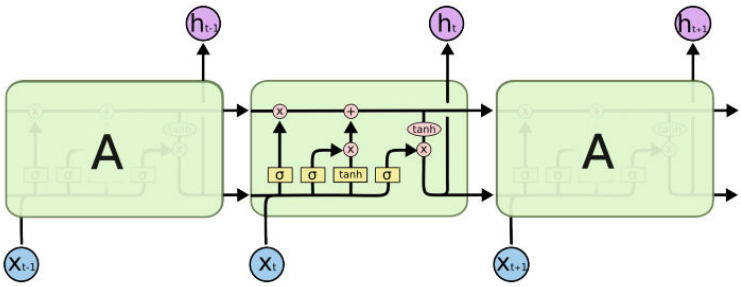


Figure 2.23: Diagram of an LSTM.

2.2. Facial Expression Analysis

The process of facial expression analysis entails the automated analysis and recognition of facial motions and feature changes via visual information [2]. This encompasses the detection and classification of facial expressions, which serve as visible indicators of an individual's internal emotional states, intentions, or social communications. The analysis extends beyond the identification of basic emotions to encompass the recognition of paralinguistic communications and subtle changes in expression that may not be immediately associated with emotions.

The architecture of facial expression analysis systems typically encompasses three primary stages: face acquisition, facial data extraction and representation, and facial expression recognition [1]. These systems developed to address the challenges inherent in naturalistic settings, including head movement, occlusion, fluctuating lighting conditions, and subtle facial movements that are prevalent in spontaneous behavior.

The study of facial expression analysis has been an enduring subject of research since the time of Charles Darwin [49] and has witnessed substantial advancements with the emergence of technologies such as image analysis, pattern recognition, and deep learning [50]. This field finds application in a variety of disciplines, including security, medicine, social sciences, marketing, and human-machine interaction, with the aim of enhancing the capabilities of systems in these areas by equipping them with the capacity to accurately and in real-time interpret human facial expressions [51].

2.2.1. Structure of Facial Expression Analysis

The fundamental architecture of facial expression analysis systems comprises three primary stages: face acquisition, extraction and representation of facial data, and facial expression classification, as seen in Figure 2.24 below.

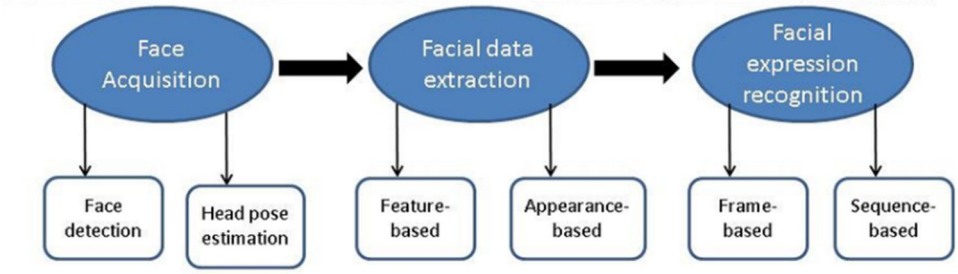


Figure 2.24: Facial expression analysis structure.

2.2.1.2. Face Acquisition

This stage represents the initial step in the process, wherein the system identifies and pinpoints the presence of a face within an image or video sequence. This may entail detecting the face in each frame of the sequence or detecting it in the first frame and then tracking it throughout the remainder of the sequence. More sophisticated systems may incorporate head finders, head tracking, and pose estimation to effectively manage significant head motion.

It is crucial for face acquisition methods to detect both frontal and non-frontal faces in any scene. Therefore, the development of face detection techniques that can accurately identify frontal and non-frontal faces across a wide range of environmental conditions is critical area of research in the field of computer vision [52]. To address out-of-plane head motion, various techniques can be implemented, such as face detection, 2D or 3D face tracking, and head-pose detection as shown in Figure 2.25. Non-frontal faces can be warped or normalized to the frontal view for expression analysis.

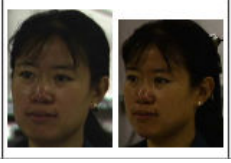
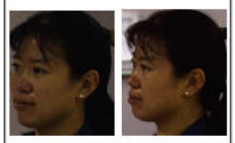
Poses	Frontal or near frontal	Side view or profile	Others
Definitions	Both eyes and lip corners are visible	One eye or one lip corner is occluded	Not enough facial features
Examples	 		 

Figure 2.25: Head pose class examples.

2.2.1.3. Facial Feature Extraction

Upon acquiring a face, the subsequent step involves the extraction and representation of pertinent facial data for expression analysis. This may include the identification of facial landmarks, the analysis of facial muscle movements, and the capture of the texture and appearance of the face. A variety of techniques, such as Gabor feature extraction and neural networks, can be employed during this stage.

Extracting facial features, which are typically classified into geometric and appearance-based categories as shown in Figure 2.26. Geometric features describe the shape and position of facial components, such as the eyes, mouth, eyebrows, and nose, and are represented by a feature vector formed from the extracted points of these components [53]. In contrast, appearance features capture the appearance of the face, such as wrinkles and furrows, and can be obtained from either the entire face or specific regions of a facial image. A face recognition system can utilize either geometric features, appearance features, or both to recognize facial expressions.

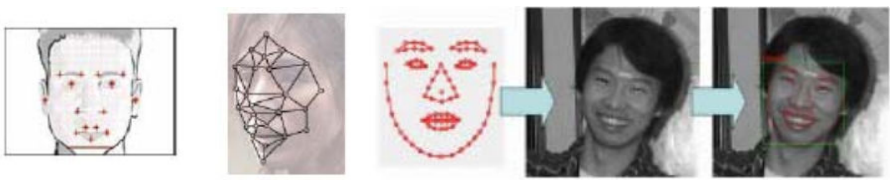


Figure 2.26: Geometric and appearance-based facial feature extraction.

2.2.1.4. Facial Expression Classification

The final phase entails categorizing the extracted facial data into one of the following expression categories: happiness, sadness, surprise, anger, fear, disgust, and neutral. This is achieved through the utilization of classifiers, such as neural networks, support vector machines, or other machine learning techniques, to identify and interpret the expressions.

In recent years, various classifiers have been implemented for facial expression recognition, including neural networks (NN), support vector machines (SVM), linear discriminant analysis (LDA), K-nearest neighbors, multinomial logistic ridge regression (MLR), hidden Markov models (HMM), and tree-augmented naive bayes. Some systems rely solely on rule-based classifications based on the definition of facial actions [37]. Frame-based recognition methods typically utilize only the current frame, sometimes in conjunction with a reference image (often a neutral face image), to identify expressions within the frame. By contrast, sequence-based recognition methods take advantage of the temporal information of sequences to recognize expressions across multiple frames. For systems that employed multiple classifiers, the highest performance in person-independent tests was selected [54]. The selection of the highest performance in person-independent tests was performed for systems that employed multiple classifiers. This process ensured that the chosen system provided the best possible accuracy and improved overall performance. Additionally, using person-independent tests allowed for a more objective evaluation of the system's performance as it eliminated the influence of individual biases and variations. By selecting the top-performing system in this manner, researchers and developers can be confident in the system's ability to provide reliable and accurate results.

An optimal system can execute these processes instantaneously and in real time, demonstrating resilience against fluctuations in age, gender, ethnicity, ambient illumination, head movements, obstructions, image resolution, and facial expressions of varying intensities. Additionally, it would possess the capacity to identify a diverse array of expressions, such as those exhibiting disparate intensities and those that are asymmetrical or unplanned. Many FER systems aim to recognize a limited set of prototypical emotional expressions, as depicted in Figure 2.27 (disgust, fear, joy, surprise, sadness, anger).

Figure 2.27: Emotion-specified facial expressions. [55]

This approach may have originated from the work of Darwin [49], Ekman and Friesen [51], and Izard et al. [53], who proposed that emotions are associated with corresponding prototypical facial expressions. However, in everyday life, these prototypical expressions are not as common. Instead, emotions are often conveyed through subtle changes in one or a few discrete facial features, such as tightening of the lips in anger or obliquely lowering the lip corners in sadness [52]. Changes in isolated features, particularly in the eyebrows or eyelids, are typical of paralinguistic displays, for instance, raising the brow signals greeting [55]. To capture the subtlety of human emotion and paralinguistic communication, it is necessary to recognize fine-grained changes in facial expressions automatically.

Extensive diversity and intricate disparities in facial expressions are noteworthy and have significant implications. These disparities encompass variations in facial plasticity, morphology, intensity of expression, and expression rate, which are not only well documented but also critical for individual identification. Such disparities can serve as reliable biometric markers to enhance the accuracy of facial recognition algorithms. Given the individual variability in expressiveness, particularly in cases of facial nerve or central nervous system damage, this is crucial. Developing robust facial expression analysis algorithms mandates the inclusion of a diverse sample of individuals from various ethnic backgrounds, ages, and sexes as well as those with distinctive characteristics, such as facial hair, jewelry, eyeglasses, and both normal and clinically impaired individuals. The accuracy of facial expression recognition can be affected by cross-cultural variations in facial structure and expression representation, as well as inter-expression resemblance. While happiness and surprise are generally the most accurately recognizable expressions across cultures, the recognition accuracy of other expressions can vary significantly due to these factors. The development of accurate and effective facial expression recognition systems necessitates the consideration of the

multidimensional nature of individual differences, including facial morphology, expressiveness, cultural background, and potential neurological impairment.

2.2.2. Facial Action Coding System

The facial action coding system (FACS: [51]) is a human-observer-based system designed to detect subtle changes in facial features. Trained observers can manually FACS code all possible facial displays, referred to as action units (AU) as shown in Figure 28, when viewing videotaped facial behavior in slow motion.

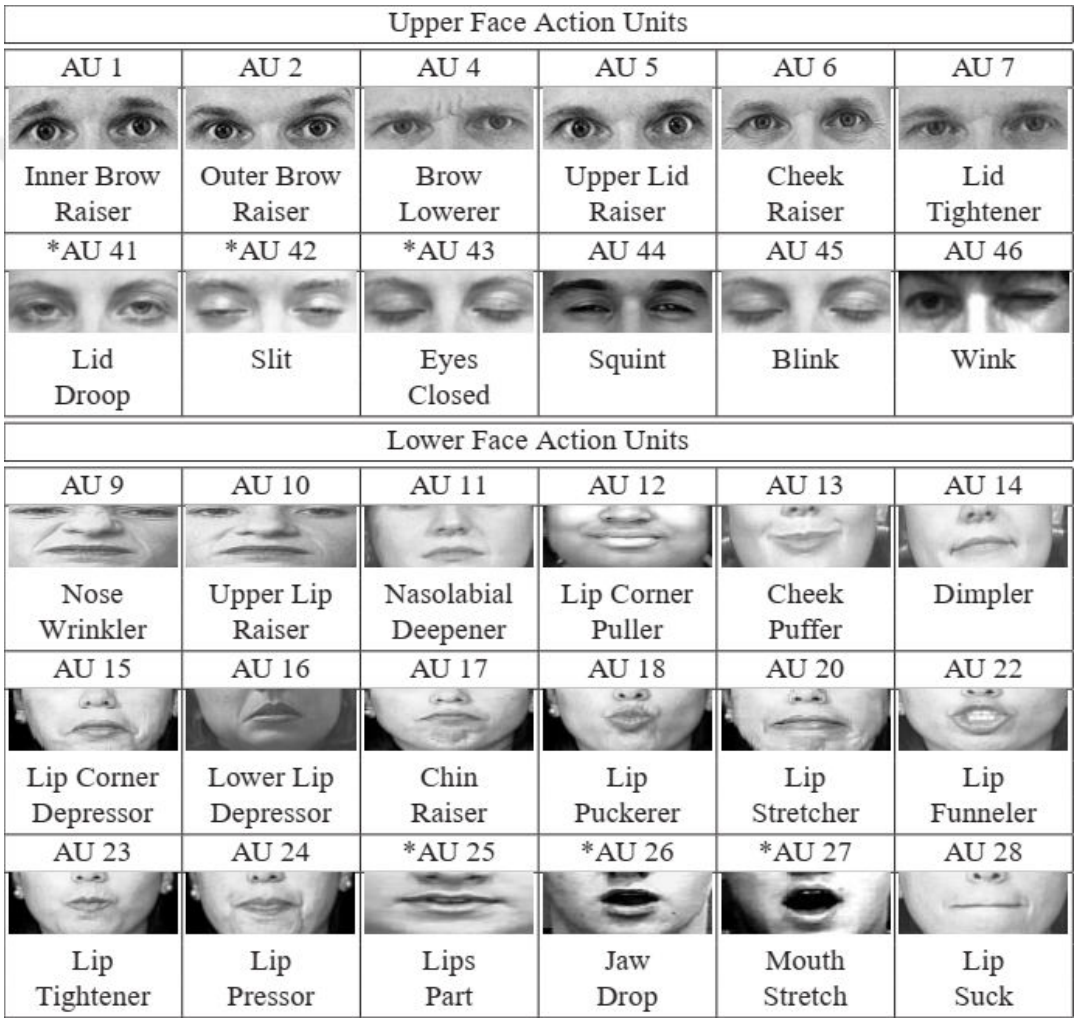


Figure 2.28: FACS action units (AU).

In facial expression analysis, it is commonly assumed that expressions begin and end with neutral faces; however, this oversimplification does not accurately reflect the complexity of facial expressions [50]. Transitions between expressions do not always pass through a neutral state, and facial expressions are composed of action units (AUs) as shown in Table 2.1, which are the fundamental actions of individual muscles or groups of facial muscles.

AU	Description
8	Lips toward
19	Tongue show
21	Neck tighten
29	Jaw thrust
30	Jaw sideways
31	Jaw clench
32	Bite lip
33	Blow
34	Puff
35	Cheek suck
36	Tongue bulge
37	Lip wipe
38	Nostril dilate
39	Nostril compress

Table 2.1: Miscellaneous Actions.

These AUs can occur in combinations as shown in Figure 2.29, and transitions may show serial dependence, meaning that one AU or a combination of AUs can lead to another without returning to a neutral state in between [56]. For example, the transition from a smile (which might involve AU 12, the lip corner puller) to a look of surprise (which could involve AU 27, the mouth stretch) does not necessarily pass through a neutral expression.

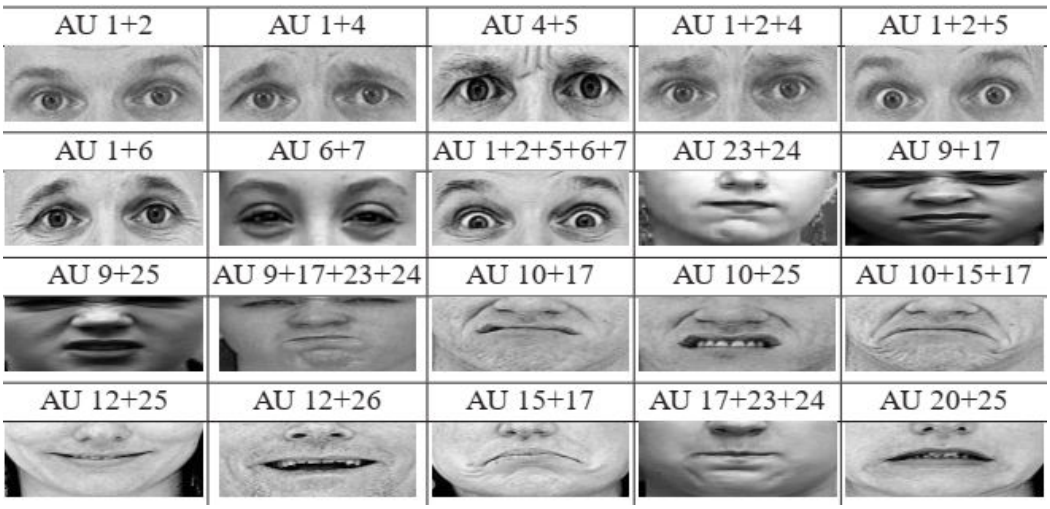


Figure 2.29: FACS action unit combinations.[57]

Additionally, the intensity of the AUs can change during these transitions by adding another layer of complexity. For instance, in the case of AU 12+27, a facial analysis

system would need to detect transitions among all three levels of mouth opening while also recognizing changes in the intensity of AU 12. Non-additive combinations of AUs, known as co-articulation effects, also occur. These interactions modify the appearance or intensity of an AU in the presence of another. For example, AU 12+15 often occurs during embarrassment [57]. To create more accurate and nuanced facial expression analysis systems, it is crucial to understand these transitions because they must be able to recognize and interpret the fluid and dynamic nature of facial expressions.

3. ETHICS OF ARTIFICIAL INTELLIGENCE

3.1. Definition of Ethics

Ethics, commonly known as moral philosophy, is a specialized area of philosophy that encompasses the development, endorsement, and presentation of moral principles governing appropriate and inappropriate conduct. Principle of ethics as shown in Figure 3.1, is defining right and wrong, proper, and improper behaviors within a specific society. These principles encompass an individual's rights and obligations, moral values, societal benefits, and sense of fairness. Ethics can differ significantly between cultures and groups, as what is accepted in one country may not be accepted in another. Moreover, ethical standards have changed over time, and what was previously considered acceptable and deemed unacceptable.

Core Ethical Values	Supporting Ethical Principles
Trustworthiness	truthfulness, sincerity, candor, integrity, promise keeping, loyalty, honesty
Respect	respect, autonomy, courtesy, self-determination
Responsibility	responsibility, diligence, continuous improvement, self-restraint
Fairness	justice, fairness, impartiality, equity
Caring	caring, kindness, compassion
Citizenship	citizenship, philanthropy, voting

Figure 3.1: Ethical values and principles.

3.2. Ethical AI

Ethical AI is guided by well-established ethical principles, including the preservation of individual rights, privacy, non-discrimination, and non-manipulation. It is critical for AI developers to consider ethical considerations when determining suitable and unsuitable applications of AI technology [7]. Organizations that employ ethical AI have formal policies and structured review processes in place to ensure that these principles are upheld. Ethical AI extends beyond the scope of what is legally permissible. Although legal restrictions on AI use establish a base level of acceptability, ethical AI encompasses policies that surpass these requirements to uphold fundamental human values. AI

algorithms that manipulate individuals, particularly teenagers, into partaking in self-harming conduct may be lawful; however, they do not embody ethical AI.

Artificial intelligence has the potential to be utilized for both beneficial and detrimental purposes [8]. The benefits of ethical and responsible AI are substantial and significant. The implementation of AI can aid organizations in becoming more efficient, producing products with a lower environmental impact, reducing negative environmental effects, improving public safety, and advancing human health. However, if utilized unethically for activities such as spreading false information, deception, exploitation of human beings, or political oppression, AI can lead to severe and far-reaching negative consequences for individuals, the environment, and society [11].

Laws and regulations typically do not provide adequate assurance for the ethical application of AI. It is the duty of individuals, organizations, developers, and providers of AI tools and technologies to embrace ethical AI principles [10]. It is crucial that AI users and suppliers actively work towards using AI in an ethical manner. Simply making statements is insufficient; strict policies must be established and enforced to ensure ethical AI practices.

The objective for AI is to exhibit desirable behavior or, at the very least, to avoid engaging in negative traits that have been identified. A comprehensive set of guidelines is essential for ensuring the development and deployment of ethical AI [9]. However, it is important to recognize that no single ethical truth applies universally.

Artificial intelligence (AI) has permeated various aspects of our daily lives, manifesting in diverse forms, including targeted advertisements during online browsing, autonomous features in vehicles, and job application screening algorithms. Media regularly covers the latest developments and applications of AI, often with considerable optimism. The ongoing digitalization of the world has substantial influence. In response to the remote learning policies implemented during the pandemic, a school in Hong Kong has integrated AI technology to analyze students' emotions from their video footage while they learn [14]. The purpose of this analysis was to evaluate the students' comprehension of the material. AI systems are now capable of performing intricate tasks such as driving

autonomously in real-life traffic, which involves a wide range of interpretation and decision-making skills.

Owing to their ability to process vast amounts of information that surpass human capacity, AI-based systems are often expected to make better decisions eventually. This expectation has already been realized in certain fields, such as cancer detection, where AI outperforms an average specialist. As this list continues to grow, our reliance on AI to make decisions has also increased. The determination of what is considered "better" by AI depends on its speed, accuracy, or optimal performance, as evaluated by specific criteria [6]. The foremost goal of AI is to accomplish or improve specific tasks, such as transporting a vehicle from one location to another. Although constraints can be implemented to address potential issues, such as avoiding harm to pedestrians during a journey, these limitations only address specific concerns. As AI becomes more complex and autonomously navigates a variety of unforeseen situations, the most effective way to ensure that it refrains from engaging in harmful actions is to imbue it with a set of universal moral principles and values. The issue of artificial intelligence (AI) causing disruption and attempts to eliminate humanity has been a recurring theme for some time. Although the prospect of AI taking over is not immediate, some of these concerns are becoming a reality, as AI is increasingly being applied in financial systems, law enforcement, autonomous vehicles, and military technology.

Artificial intelligence (AI) has generated significant ethical dilemmas in the 21st century. AI offers advantages such as increased efficiency, reduced human error in medical diagnosis, and utilization of robots in hazardous situations, such as securing a nuclear plant after an incident. However, AI also presents ethical conundrums, including algorithmic bias, digital divide, and severe health and safety issues. The AI ethics industry has grown, encompassing a diverse array of stakeholders. Nevertheless, AI ethics is not a novel concept [7]. The notion of AI has existed for nearly 70 years and ethical concerns have been raised since the mid-twentieth century. This debate has gained momentum because of concerns regarding the consequences of superior algorithms, increasing computing resources, and growing amount of data available for analysis.

AI ethicists face challenges in comprehending ethical issues because of the lack of transparency in AI development and deployment. However, their role is not to program

the systems themselves but to understand the ethical implications of AI design, development, and deployment. This entails understanding concepts, such as supervised and unsupervised learning, dataset labeling, and user consent. Although AI developers and ethicists may not fully understand the intricacies of advanced algorithms, it is essential for AI ethicists to have a sufficient grasp of the technology to identify key ethical questions that require addressing. Although AI is not entirely transparent, ethical concerns must be considered.

3.2.1. Ethical AI Frameworks

Numerous organizations and groups across various nations have issued extensive AI principles developed by substantial panels of experts, each providing a particular perspective. These principles are often labeled with brand names that reflect the specific methodology adopted by the group [11]. The sheer volume of these principles can be overwhelming, and they are occasionally vaguely defined and confused, with names such as Trustworthy AI, Ethical AI, Transparent AI, Human-centered AI, Safe AI, and others. Nevertheless, the principles themselves are the most critical, and it is evident that many of them are shared among different frameworks [14].

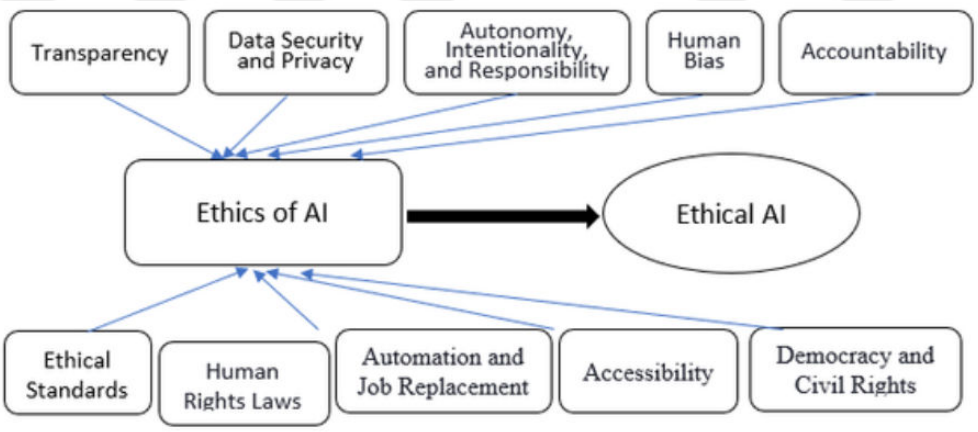


Figure 3.2: Framework for AI ethics.

The ethical principal framework for AI as shown in Figure 3.2 above, is described by a variety of terms that reflect the general mindset and spirit of the principles as follows:

- **Ethical AI** is designed to adhere to ethical principles and to ensure compliance with moral standards.

- **Trustworthy AI** operates in a manner that is both lawful, ethical, and dependable, thereby gaining the trust of humans.
- **Explainable AI** allows stakeholders to comprehend the functioning of an AI in non-technical terms.
- **Interpretable AI** allows stakeholders to not only understand but also scrutinize the decision-making process of AI.
- **Meaningful AI** is environmentally friendly and does not exacerbate exclusion or inequality but is also explainable.
- **Transparent AI** provides some level of accessibility to data or algorithms, thereby promoting openness and accountability.
- **Responsible AI** considers societal values, morals, and ethical principles.
- **Human-centered AI** places human value at the forefront of AI system development, deployment, use, and monitoring.
- **Beneficial AI** not only mitigates risks, but also actively contributes positively to society.

3.2.2. AI Bias

Bias constitutes a systemic deviation in decision-making procedures that culminates in unequal results. Within the context of artificial intelligence, bias can originate from various sources, including data acquisition, algorithm formulation, and human comprehension [69]. Machine learning models, which are a type of artificial intelligence system, possess the capacity to learn and perpetuate the prejudices present in the data utilized for their training, ultimately resulting in unfair or discriminatory outcomes. Identifying and rectifying bias in AI is of paramount importance to guarantee that these systems are fair for all users [34]. This section examines the origins, repercussions, and methods of mitigating bias in AI at greater depths.

Bias in AI may originate from multiple stages of the machine learning process, including data acquisition, algorithm formulation, and user engagement. Data bias refers to a situation in which machine learning models generate biased outputs owing to unrepresentative or incomplete data. This may occur when data are sourced from biased sources, when they lack vital information, contain errors, or are incomplete. However, algorithmic bias occurs when the algorithms used in machine learning models possess inherent biases that are reflected in their outputs. This may occur when algorithms are

developed based on biased assumptions or when biased criteria are used to make decisions [37]. Finally, user bias arises when individuals using AI systems consciously or unconsciously introduce their own biases or prejudices into the system. This may occur when users provide biased training data, or when they interact with the system in ways that exhibit biases [60].

Although multiple solutions have been suggested to tackle these biases, the most effective ones include dataset expansion, the development of bias-conscious algorithms, and the incorporation of user feedback mechanisms. Dataset expansion involves incorporating diverse data into training datasets to improve representativeness and reduce bias. On the other hand, bias-conscious algorithms are designed to consider various types of biases and limit their impact on the system's output. Finally, user feedback mechanisms seek input from users to identify and rectify the biases in the system [62]. These strategies are crucial for addressing biases in AI systems and ensuring fairness in their outputs.

Table 3.1 below, serves to illustrate the diverse array of biases that can have significant negative consequences within AI systems. This underscores the critical need for extensive evaluation and the application of effective mitigation strategies to address these biases which are occurring in several types.

Bias Type	**Description**	**Example**
Sampling Bias	When training data does not accurately reflect the population, it is intended to serve, it can result in suboptimal performance and biased predictions for specific groups.	Facial recognition algorithms, particularly those primarily trained on white individuals, tend to exhibit subpar performance when processing the faces of individuals from other racial backgrounds.

Algorithmic Bias	The results of the creation and execution of the algorithm could potentially give preference to specific attributes and result in unfair outcomes.	The consequences of the algorithm's development and execution might encompass the favoritism of specific characteristics, culminating in disparate outcomes.
Representation Bias	When a dataset fails to accurately reflect the population that it is designed to represent, it can result in inaccurate predictions.	A deficient medical dataset that fails to adequately represent women can result in a less precise diagnosis for female patients.
Confirmation Bias	Emergence is the phenomenon that occurs when an AI system is put into use and verifies the pre-existing biases or convictions that its developers or users already possess.	An AI system that assesses the success of job candidates based on the biases of the hiring manager.
Measurement Bias	When data collection or measurement often neglects certain demographics, either by excluding them entirely or underrepresenting them inadequately.	The survey, which aimed to gather many responses from urban residents, unintentionally led to a lack of representation for rural perspectives.

Interaction Bias	Occurs when an AI system displays biased behavior towards humans, resulting in unfair treatment.	An AI system that displays disparate responses based on the gender of the user, thereby reinforcing biased communication.
Generative Bias	Generative AI models can suffer from a generative bias, which produces an unequal representation of certain attributes or perspectives in the output, leading to unbalanced content.	Models, trained on literature from Western authors, may neglect diverse cultures and idioms, while image generation models with limited diversity in datasets struggle to accurately represent various ethnicities.

Table 3.1: Types of AI biases.

3.2.2.1. Challenges of Datasets in AI Bias

AI-driven methods typically depend on substantial datasets to support machine learning operations. Experts and practitioners often opt for readily available data to identify the most suitable dataset, which may result in data being sourced from non-representative channels, such as online surveys or social media platforms. Such datasets may exhibit biases or lack a comprehensive representation of the target population and are frequently utilized to train machine learning models deployed in diverse sociotechnical contexts, potentially marginalizing certain societal groups. Availability biases can also occur when datasets that are easily accessible but not fully representative of the target population are constantly employed as training data, potentially resulting in under-representation of disadvantaged populations, including indigenous communities, women, and individuals with disabilities. In addition, issues can emerge in natural language processing (NLP) applications when datasets do not align with real-world situations, leading to discriminatory outcomes and performance disparities [67].

AI systems depend heavily on substantial historical data, which are often refined using machine-learning models. However, these datasets may be unbalanced, leading to biased outcomes and representativity issues when sampled. Simple random sampling, the most widely used method in statistical surveys, requires that the probability of sample extraction be known and not zero and that each element and each combination of elements have an equal probability of being selected [72]. Biased samples yield biased estimates. Therefore, statistical sampling is crucial; however, in the era of Big Data, much of the data used today is not generated using probabilistic sampling methods. Instead, it is often obtained from third parties or through opportunistic methods made possible by digital technology. These non-probabilistic methods do not offer equal opportunities for every unit of the population to be included in the sample, which means that certain groups or individuals are more likely to be chosen, whereas others are less likely [77]. Representativity is the quality of the sampling process that is inherently random; therefore, it is essential to maintain control over this aspect when using non-probabilistic samples.

Implementing demographic or statistical parity solutions can be beneficial when there is no intention of differentiating a protected group that would otherwise face penalties. However, the effectiveness of these solutions depends on the characteristics and objectives of the data being analyzed. For example, in a study that considers individual income as a factor, representativeness issues do not arise if high-income individuals voluntarily participate in the sample because the selection of a particular group aligns with the objectives of the study. Nevertheless, if lower-income individuals are less likely to be selected, the average sample income will exceed the overall population income.

Employing a model with biased or unfair dataset may generate prejudiced and incorrect forecasts. It is imperative to exercise prudence when applying a model trained on one dataset to another because disparities in the distribution of datasets may exacerbate the model's unfairness and inaccuracies.

4. LITERATURE REVIEW

4.1. Fairness and Bias in FER Models

As AI advances, concerns have been raised about its potential consequences owing to significant strides made in affective computing and extensive research conducted on facial expressions and deep learning. In a study published in [60], the authors discussed gender bias in the context of responsible AI and emphasized the need to address race as another bias. To better understand the relationship between race and affective computing, psychological research has conducted numerous studies on the correlations between culture, ethnicity, gender, and emotions [61-65]. For instance, one study found that individuals within the same race could accurately identify their emotions from either the same or different races [67]. However, misclassification occurred. Although AI has the potential to surpass human performance due to its ability to learn faster and more accurately, it is essential to remember that machines can still experience overfitting. Therefore, all necessary precautions must be taken to prevent this from happening.

Fairness is a central ethical principle for AI systems, as outlined in trustworthy AI guidelines [68]. Along with respect for human autonomy, prevention of harm, and explicability, it is one of the four key principles of AI ethics [69]. However, the term fairness can be ambiguous, leading to debates and discussions. Bias in machine learning can be viewed from various perspectives, including biased sampling, inappropriate feature selection, biased labeling, and more. The AI industry is actively working to identify and mitigate the different types of biases in machine learning to address the risks associated with AI. Several studies have explored the issue of unfairness in machine learning (ML) systems. Furthermore, researchers have examined the root causes of bias in ML models. Evaluation metrics that consider the social context are employed to assess the fairness of these models. It is crucial that these evaluation metrics are utilized to ensure that AI models are developed and deployed responsibly and ethically without perpetuating biases or discrimination in the social context.

A paper published in [68] discussed the prerequisites and obstacles for creating responsible AI, emphasizing the development of beneficial social applications, and reducing the hazards of AI systems. The authors also highlighted instances of bias arising from a lack of transparency, comprehensibility, and biased training data that are difficult

to detect and rectify. In [69], practical advice was provided on constructing responsible AI using a design methodology, stressing the significance of using technical tools such as Facebook's Fairness Flow, IBM's Fairness 360 toolkit, and Accenture's AI Fairness tool to identify and mitigate bias in sensitive datasets. In [69], researchers found inadequacies in conventional AI definitions as rational agents, emphasizing the need for a broader range of ethical principles to create more socially accountable AI agents. They discovered a flaw in a credit card dataset that could result in racial bias, even though it did not contain protected attributes such as race. Other personal information can still be used in a discriminatory manner when analyzing datasets.

The shortcomings of deep learning in providing clear input features and explaining decisions have resulted in the discovery of various biases. In addition, 20 distinct biases were documented in [37]. For example, Representation Bias [33] impacts well-known datasets such as ImageNet [8] and OpenImages [7], owing to a disproportionately represented demographic group in the data collection process 11 [69]. Another prevalent form of bias is Population Bias [34], which occurs when the statistics, demographics, representatives, and user characteristics of a dataset or platform differ from those of the original target population [14]. In essence, representation bias is rooted in the way data are collected from the population, whereas population bias may be concealed within the population itself. Discrimination against underrepresented groups in AI for facial emotion recognition has been identified, revealing racial bias. This is evidenced by the detection of bias in Microsoft's Face and Face++ API's as well as studies demonstrating a relationship between gender and the presence of smiles (indicating positive emotions) in the CelebA dataset.

In a study conducted by the authors [73], three different scenarios were evaluated for ethnicity classification: Black and Caucasian individuals, Chinese and non-Chinese individuals, and Han, Uyghurs, and non-Chinese individuals. The authors utilized deep convolutional neural networks (DCNN) to simultaneously extract and classify facial features. The process involved detecting, aligning, normalizing, and cropping the face, which was trained using a DCNN model. The authors' method was compared to other techniques, including Biologically Inspired Features (BIF), Kernel Class-dependent Feature Analysis (KCFA), and Local Binary Pattern (LBP), and was found to outperform them owing to its training with large-scale datasets such as MORPH-II, CASIA-PEAL,

and CASIA-WebFace. Similarly, the authors of [74] trained their model using different datasets, including CK+, JAFFE, and BU-3DFE, to predict facial emotions. Other similar studies include [75], where the authors analyzed Western Caucasian and East Asian facial expressions of emotions based on visual representations and cross-cultural FER, and [76], where the authors proposed a joint deep learning approach called racial identity-aware deep convolutional neural network to recognize multicultural facial expressions and also contributed to the field of cross-cultural emotion recognition, where the authors proposed a novel method for facial expression analysis using machine learning techniques.

To minimize racial bias in Federal AI, Livingston, M. [71] proposed three recommendations: fostering diversity among AI developers, conducting AI impact evaluations, and implementing procedures for employees to challenge digital decision-making processes. According to Livingston [71] these strategies are essential to ensure that AI systems are fair and equitable for all users. By promoting diversity in the development of AI systems, it is possible to incorporate a broader range of perspectives and experiences, thereby reducing the potential for racial bias. In addition, conducting AI impact evaluations can help to identify and address any biases that may exist within AI systems. Finally, implementing procedures for employees to challenge digital decision-making processes can promote transparency and accountability in the development and deployment of AI systems, further minimizing the potential for racial bias. Linked bias is the cause of underrepresentation of affected populations in the field of technologists, leading to a lack of racial diversity. Therefore, AI models perform better when trained on datasets with balanced ratios of Caucasian and Black individuals [72]. Researchers have provided further evidence for this perspective in [71]. These studies emphasize the importance of considering the specific context in which a person decides as well as the role of emotions and social influence in shaping their choices. Additionally, these studies suggest that a more holistic decision-making approach can lead to better outcomes and greater satisfaction with choices made. Overall, the evidence presented in [72] further supports the argument for considering the complex interplay of cognitive, emotional, and social factors in decision-making processes.

According to a study conducted by the authors in [62], a thorough investigation was conducted to determine what can be learned from facial features. This study was based on interdisciplinary expertise and provided an analytical foundation for understanding

race. This knowledge was then utilized to create a comprehensive framework that examined the intricacies of races from numerous perspectives. The authors also emphasized the importance of excluding hair from facial recognition models to ensure that they are unbiased.

Racial bias is an insidious issue that affects various domains, including facial recognition technology. Studies have demonstrated that image classification models can incorrectly classify Black men as "primates," as reported by Buolamwini and Gebru [84]. This problem gained widespread attention following the publication of their research [84]. For instance, a popular science magazine sparked a nationwide conversation on the topic. In response, Microsoft, IBM, and Face++ have updated their APIs, resulting in improved performance metrics [70]. IBM removed facial detection from its API in September 2019. In 2019, San Francisco became the first US city to ban the use of face recognition technology because of concerns regarding bias in the technology [11]. Other cities have since followed suit. Since then, a substantial amount of research has been conducted to address and reduce biases in facial recognition technologies, leading to advancements in this field. To alleviate racial bias, researchers have primarily adhered to three guidelines: **a)** balancing datasets, **b)** designing model selection and loss, and **c)** erasing sensitive features from facial representations utilized for identification. This study investigates the impact of a specific blinding method on the removal of sensitive features from face matching techniques in accordance with the third guideline. Recent endeavors to address biases in machine learning and facial recognition have been summarized in systematic reviews [,]. These reviews offer a comprehensive overview of the various approaches and methods utilized to address biases in machine learning and facial recognition as well as their effectiveness in reducing biases and enhancing fairness in these systems.

In most studies, human races are typically classified into eight categories: African, African American, Caucasian, East Asian, Native American/American Indian, Pacific Islander, Asian Indian, and Hispanic/Latin. The study in [66] also referenced a variety of commonly used race-related facial databases, such as CAS-PEAL, IFDB, Texas 3DFRD, KFDB, JAFEE, CAFE, FGRC 2.0, CMU DB, BU-3DFE, FERET, CohnKanade, Asian PFOI, HAJJ and UMRAH, JACFEE, CUN, Indian DB, FEI, and EGA.

A comprehensive study, referred to as [107], was recently conducted to examine bias and fairness in FER, with a focus on three specific biases: age, race, and gender. The CelebA [58] and RAF-DB [86] databases were utilized to conduct the experiments. Despite dividing the test data into male and female groups, conclusions were not drawn regarding the per-class differences in accuracy. Instead, the study found that the model displayed a bias towards the female group, resulting in an overall accuracy score of 3%. Additionally, the study revealed that gender bias was less significant than age and race biases. Addressing racial biases in face recognition technology begins by addressing the biases in the data used to train the algorithm. This includes addressing measurement error, which refers to systematic errors in the measurement of variables for specific groups, and representation bias, which occurs when not everyone has the same probability of being included in the dataset, resulting in a lack of diversity in the training data that does not reflect the real world. To address the issue of dataset imbalance, researchers have proposed using synthetic data to create racially balanced datasets. For instance, Kortylewski et al. [87] suggested using synthetic data to balance the dataset, whereas Robinson et al. introduced a racially balanced dataset and demonstrated how adapting decision thresholds for each race can reduce performance gaps in face recognition, leading to improved fairness and reduced disparities in outcomes among different racial groups in the field of face recognition. Similarly, recent studies have shown that incorporating cultural and ethnic diversity into facial recognition algorithms can significantly improve accuracy and fairness. By considering the varying physical features and expressions of individuals from diverse cultural backgrounds, these algorithms can better recognize and accurately identify individuals across different racial and ethnic groups. This approach has demonstrated potential in reducing racial and ethnic bias in facial recognition systems, promoting fairness and equality in law enforcement, and other applications.

Deep learning fairness improvement methods are typically categorized into pre-processing, in-processing, and post-processing techniques, depending on the stage at which they are applied (Mehrabi et al. [88], Oneto & Chiappa [89], Mehrabi et al. [90]). Furthermore, AI fairness toolkits that utilize these methods have been developed [92]. Researchers have also identified biases in expression labeling within datasets, which are influenced by the impact of race on emotion perceptions, as contributors to unfairness [91]. To address this issue, AI fairness toolkits can be used to address biased expression

labeling within datasets, leading to unfair outcomes [92]. Additionally, studies have found that the impact of race on emotion perception contributes to these biases in expression labeling [37]. However, Chen and Joo's study [15] (2021a) did not report any systematic labeling biases for races and attributed the absence of imbalanced racial representations in the (2021a) dataset.

Researchers have pursued various strategies to eliminate racial bias in face recognition, such as adjusting the model choices and targets. One method involves customizing the selection of face samples for training based on the data distribution and model bias [73]. Another strategy is to transfer knowledge from the source domain to the target domain by learning the facial features with adequate generalization across different races [74]. Furthermore, researchers have employed Generative Adversarial Networks (GANs) [93] to develop blind models and minimize the correlation between sensitive features and facial attributes in face recognition [94]. Adeli et al. proposed an adversarial loss to decrease the correlation between model representations and sensitive information (races), and the statistical dependency of the learned features and a source of bias (racial group) [46]. To assess the effectiveness of their proposed method, Adeli et al. conducted experiments on a diverse range of datasets and found that their approach significantly reduced the correlation between model representations and sensitive racial information, as well as the statistical dependency of the learned features and a source of racial bias.

The mitigation of racial bias in facial emotion recognition has been accomplished through the implementation of certain effective methods [7]. However, the success of these methods relies heavily on the datasets and models utilized. The incorporation of highly subjective labels and high-dimensional inputs in emotion recognition presents a formidable challenge for achieving comprehensive fairness in this domain and similar areas, such as complex and heterogeneous data streams. Additionally, the lack of standardization in emotion recognition datasets impedes researchers' ability to compare and integrate data from various sources, thereby hindering the development of more accurate and fair emotion recognition models.

4.2. Deep Learning Based FER Approaches

Researchers in the field of facial emotion recognition using deep learning often utilize convolutional neural networks and recurrent neural networks. Khorrami et al. [95] found

that CNNs are particularly adept at accurately detecting facial action units (FAUs), resulting in promising classification performance on the CK+ dataset. Lopes et al. [74] proposed a combination of CNNs and image preprocessing techniques, such as cropping and normalization, to reduce the number of convolutional layers and minimize the need for extensive training data. This led to an improved overall accuracy and computational efficiency for the CK+, JAFFE, and BU3DFE datasets. Agrawal and Mittal (2020) [31] introduced a novel CNN model that investigated the impact of kernel size and the number of filters on the final classification accuracy, resulting in further performance enhancement. Recent studies have explored the potential of integrating generative adversarial networks (GANs) into CNNs for data augmentation and training [93-94]. These studies have demonstrated the effectiveness of GANs in improving the performance of CNNs in various applications such as image classification, object detection, and segmentation.

Deep learning models play a crucial role in every stage of FER [96]. To effectively train deep neural networks for FER, diverse datasets spanning a range of labels and data types are necessary [97-100]. Li and Deng (2018) [16] presented an overview of widely used datasets specifically designed for deep emotion recognition. Li and Deng's (2018) [2] review of popular datasets developed for deep emotion recognition underscores the growing interest in enhancing the accuracy and robustness of models for detecting and analyzing human emotions. These datasets offer valuable resources for researchers and practitioners alike, enabling them to assess and optimize their models in a more realistic and challenging environment. As the field continues to advance, it is likely that even more sophisticated datasets will be developed, further enhancing the capabilities of deep emotion-recognition technology.

Zhu et al. (2017) [101] proposed a novel approach to facial emotion recognition by incorporating convolutional neural networks with recurrent neural networks in a bidirectional architecture (2017). RNNs have also been shown to be effective in capturing dynamic facial actions within a multimodal FER, as demonstrated by Majumder et al. [102]. Despite achieving high overall accuracy, the performance of these models across different racial groups is yet to be thoroughly and systematically studied. Although multicultural FER models have shown effectiveness in improving accuracy, their impact on fairness is yet to be explicitly addressed [76]. To address this gap, it is crucial to

conduct further research on the impact of multicultural FER models on fairness and to develop strategies to ensure that these models promote fairness and accuracy simultaneously. Moreover, it is important to investigate the effectiveness of different multicultural FER models and compare their performance in terms of both fairness and accuracy to determine which model(s) are most suitable for specific applications.

5. METHODOLOGY

5.1. Database

The primary emotions used on this study: happiness, neutrality, and sadness, which represent positive, neutral, and negative valence, respectively. These emotions are among the most experienced online emotions and serve to mitigate potential biases that may emerge at any point, from data acquisition to evaluation. AffectNet [97] dataset is used to examine and analysis the racial bias.

5.1.1. AffectNet Dataset

AffectNet [97] is a prominent facial expression dataset which is the largest in the world. It was compiled from images sourced from the Internet and comprises eight emotion classes: happy, sad, anger, fear, surprise, disgust, contempt, and neutral. AffectNet contains 287,651 images in the training set and 4,000 images in the validation set, with each class (expression) consisting of 500 images. The total annotated data distribution is illustrated below in Table 5.1.

Expression	Total
Neutral	75374
Happy	134915
Sad	25959
Surprise	14590
Fear	6878
Disgust	4303
Anger	25382
Contempt	4250
None	33588
Uncertain	12145
Total	**420299**

Table 5.1: Total number of annotated images of AffectNet.

Owing to the unavailability of the test set, most studies on AffectNet employ the validation set as the test set. As mentioned before, only three emotions selected for this study—**happy**, **sad**, and **neutral**—other expression groups are excluded from the original training set. Consequently, the final training set used comprises the size of N=236,248 images, with 134,915 images classified as happy, 75,374 as neutral, and 25,959 as sad. Similarly, the validation set size is specified as N=1500.

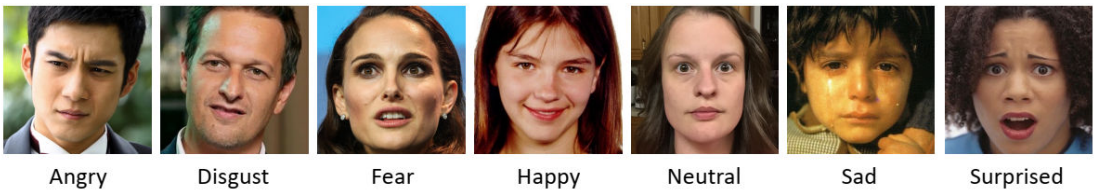


Figure 5.1: Sample images from AffectNet dataset.

On the other hand, AffectNet does not provide racial information as shown from sample images in Figure 5.1 above. Therefore, the distribution estimation is obtained using a model developed based on the FairFace [103] presented classifier in which $\hat{Y}$ is the predicted gender, Y is the true gender, A refers to demographic group and D is the set of considered race groups:

$$P(\hat{Y}=i|Y=i,A=j) = P(\hat{Y}=i|\hat{Y}=i,A=k)\,,\ i \in \{male, female\}, \forall j,k \in D$$

$$\varepsilon(\hat{Y}) = \max_{\forall j,k \in D}\left(log\frac{P(\hat{Y}=Y|A=j)}{P(\hat{Y}=Y|A=k)}\right)$$

Models as shown in Figure 5.2, trained using the FairFace's labeled race dataset [103], demonstrate greater racial balance accuracy than the other comparable datasets, which are often led to asymmetric accuracy problems on different races, the performance of models trained using the FairFace dataset is superior for non-white faces compared to other models and datasets.

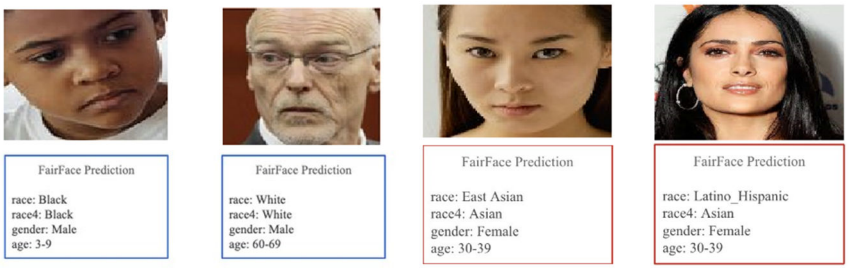


Figure 5.2: FairFace based model's classification examples.

Original weights provided by FairFace is used to predict racial information of the AffectNet's training set, model's some prediction examples can be seen from the figure above. Utilizing the FairFace-based prediction model [103], AffectNet face images [97] are categorized into four distinct groups: **White**, **Black**, **Asian**, and **Indian**. Figure 5.3 is showing the examples of the race-based classification samples of AffectNet [97].

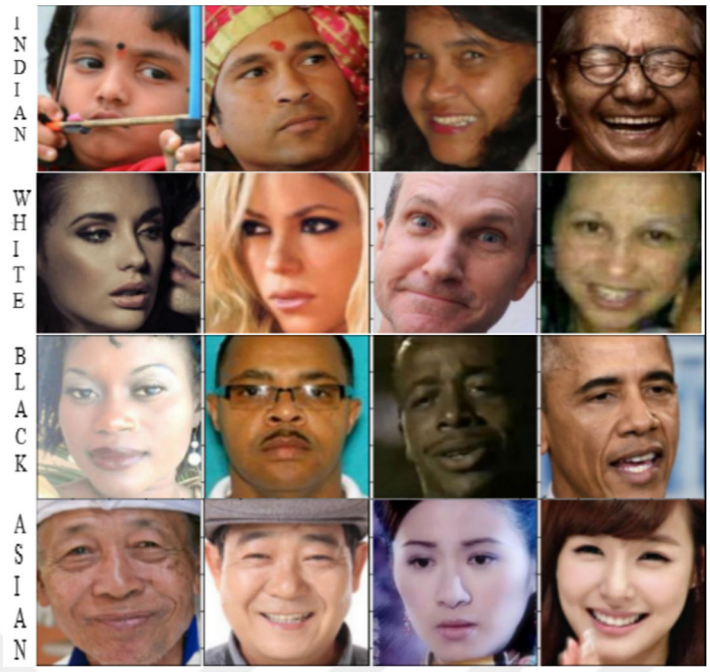


Figure 5.3: Classified image samples from the AffectNet. [97]

The following table (Table 5.2) provides the results of precise numbers of faces and race proportions gathered from FairFace model's [103] estimation on AffectNet dataset.

			Expression			Counts	
	Dataset	**Abbr.**	**Happy**	**Neutral**	**Sad**	**Total**	**Proportion %**
Race	**White**	**W**	109,368	56,330	15,638	181,336	**76,757**
	Black	**B**	8,429	8,445	2,708	19,582	**8,289**
	Asian	**A**	9,582	6,359	5,162	21,103	**8,933**
	Indian	**I**	7,536	4,240	2,451	14,227	**6,022**
Total			**134,915**	**75,374**	**25,959**	**236,248**	**100**

Table 5.2: Content and race distribution of the dataset.

As expected, the White race group constituted the largest proportion of the training set and the Indian race group is the smallest distributions for AffectNet [97], respectively comprising 76,557% and 6,022% of their respective datasets. As shown in Table 5.2, the datasets are characterized by an imbalance in the number of samples per emotion. To rectify this issue, a decision is made to create smaller, balanced class samples. Among all

the emotions, the fewest samples were 2451. Thus, to ensure uniformity in the distribution of emotions, the size of each set was reduced to 2000 images for each emotion, resulting in a total of 6000 images. This approach is adopted to prevent population bias and to investigate the consequences of having imbalanced classes in comparison to balanced emotion sample counts. Some example images from the datasets can be seen in the Figure 5.4.

Figure 5.4: Sample images from the dataset.

5.2. Frameworks

In the field of facial expression analysis, novel deep learning architectures and transfer learning methods are commonly used by researchers owing to their superior performances.

In this study, two state-of-the-art architectures, namely Deep-Emotion and Self-Cure Network, are evaluated, in addition, the mainstream models of three widely used state-of-the-art transfer learning methods, namely ResNet50 [80], InceptionV3 [81], and DenseNet121 [82] which are provided by Keras and Pythorch, are adopted to provide a proof of concept for racial bias investigation.

5.2.1. Deep-Emotion

Deep-Emotion [78] is a deep learning framework that employs an attentional convolutional network to classify the underlying emotions present in facial images. The architecture of the Deep-Emotion model is shown in Figure 5.5.

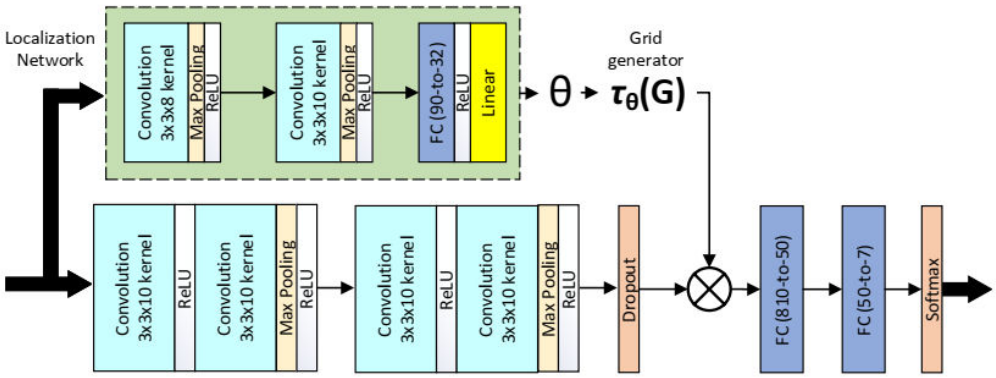


Figure 5.5: Deep-Emotion architecture.

The feature extraction component of Deep-Emotion consists of four convolutional layers, each followed by a max-pooling layer, a rectified linear unit (ReLU) activation function, a dropout layer, and two fully connected layers. The spatial transformer, or localization network, comprises two convolutional layers, each followed by max-pooling and ReLU, as well as two fully connected layers. After regressing the transformation parameters, the input is transformed into the sampling grid $T(\theta)$ to produce warped data. The spatial transformer module focuses on the most relevant parts of the image by estimating a sample over the region of interest. The loss function of Deep-Emotion is a combination of two terms, which include the classification loss (cross-entropy) and the regularization term (ℓ), totaling their values to form the outcome as seen in equation below:

$$\mathscr{L}_{overall} = \mathscr{L}_{classifier} + \lambda \left\| w_{(fc)} \right\|_2^2$$

The weight of regularization denoted by λ, is modified to determine the optimal value that leads to the most favorable results on the validation set. This is achieved by employing a combination of dropout and ℓ_2 regularization, allowing for the training of models from the ground up, even when dealing with extremely limited datasets.

5.2.2. Self-Cure Network

The Self-Cure Network (SCN) [79] is built upon the fundamentals of traditional Convolutional Neural Networks and comprises three key modules: self-attention importance weighting, ranking regularization, and relabeling, as illustrated in Figure 5.6.

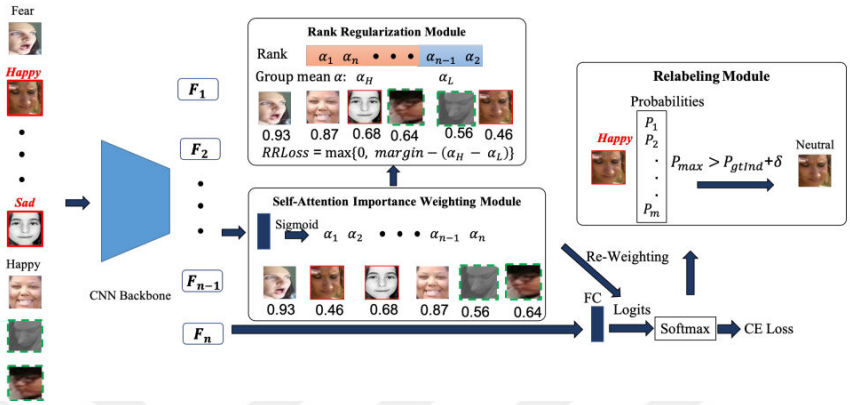


Figure 5.6: Self-Cure Network architecture.

As illustrated in the figure below, the process begins by feeding facial images into a backbone CNN for feature extraction. Subsequently, the self-attention importance-weighting module learns the sample weights from the facial features for loss weighting. Next, the rank regularization module takes the sample weights as inputs and constrains them using a ranking operation and a margin-based loss function. Finally, the relabeling module searches for reliable samples by comparing the maximum predicted probabilities with those of the given labels.

5.2.3. Transfer Learning

Transfer learning is an approach that allows pre-trained neural networks to apply the knowledge gained during their initial training to a different task. This process enhances the performance and generalization capabilities of the network, particularly when the dataset is insufficient. This technique is commonly used in facial emotion recognition tasks. The generation of a transfer learning framework model can be seen in the Figure 5.7 below.

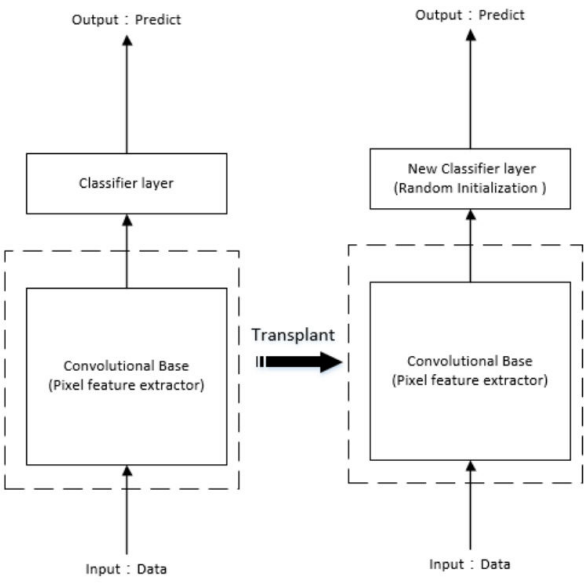


Figure 5.7: Transfer learning model generation.

In this study, the goal is to evaluate the performance of three transfer learning techniques that involve three widely used deep convolutional neural network architectures; therefore, three transfer learning models, Resnet50 [80], InceptionV3 [811], and Dense121 [82], are implemented individually by constructing separate models refined through training on the constructed database for model training.

5.2.3.1. ResNet50

Residual Network, referred to as ResNet, is a particular category of convolutional neural networks that were first introduced in 2015 [80]. The original ResNet architecture, known as ResNet-34, comprised 34 weighted layers. This design addresses the vanishing gradient problem, which typically restricts the depth of the CNNs by incorporating shortcut connections [83]. A shortcut connection bypasses certain layers, thereby transforming a regular network into a residual network. ResNet-50, for instance, includes 50 layers, featuring 48 convolutional layers, one MaxPool layer, and one average pool layer. Residual neural networks are a type of artificial neural network that are formed by stacking residual blocks to mitigate the vanishing gradient problem and improve network depth.

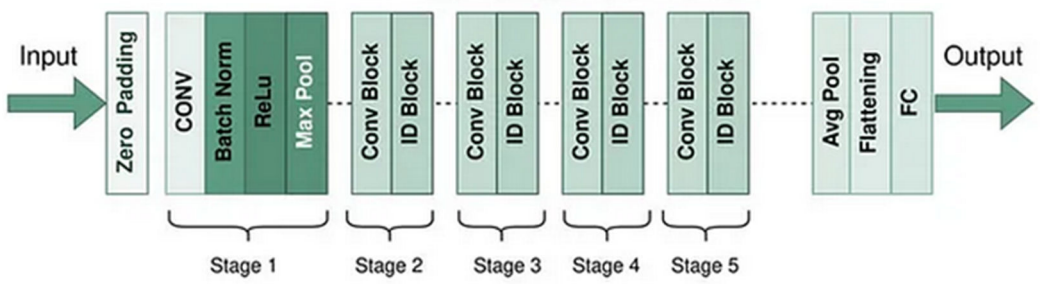


Figure 5.8: ResNet50 Architecture.

As illustrated in Figure 5.8 above, the ResNet architecture is governed by two primary design principles. Initially, the number of filters in each layer is equivalent to the dimensions of the output feature map. Second, when the feature map size is reduced by half, the number of filters is increased twice as much to maintain the time complexity of each layer.

5.2.3.2. InceptionV3

Inception V3 [81] is a deep learning model that utilizes Convolutional Neural Networks with 42 layers for image classification. This model represents an advanced version of the foundational Inception V1 model, originally introduced as GoogLeNet in 2014. Developed by a team of experts from Google, Inception v3 is designed to reduce the computational power consumption by modifying previous Inception architectures. Inceptionv3's model diagram is illustrated in the Figure 5.9 below.

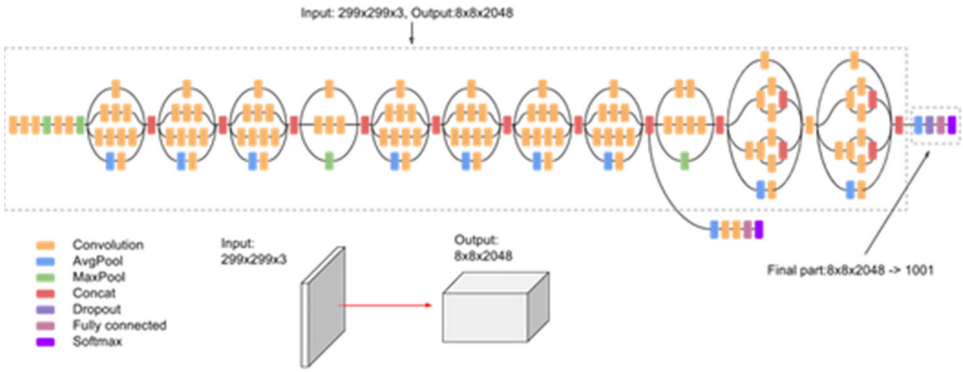


Figure 5.9: Inception v3 high-level diagram.

The Inception Network, particularly GoogLeNet/Inception v1, has been shown to be more efficient in terms of the number of parameters generated and the cost incurred (including memory and resources). The modification of an Inception Network while maintaining its

computational advantages requires careful consideration. This is because adapting Inception Networks for various applications can be challenging owing to the uncertainty in the efficiency of the new network. To address this, the Inception v3 model proposes several optimization techniques, including factorized convolutions, regularization, dimension reduction, and parallelized computations, which facilitate easier model adaptation while preserving the computational efficiency.

5.2.3.3. DenseNet121

DenseNet121[83] architecture utilizes structures called dense blocks, which has 121 layers in its architecture and incorporate dense blocks consisting of layers with a feedforward connection. The feature maps from all the previous layers in each layer of DenseNet are utilized as inputs. This feature reuse facilitates the information flow in the model and allows for visualization of the task flow in the DenseNet121 model.

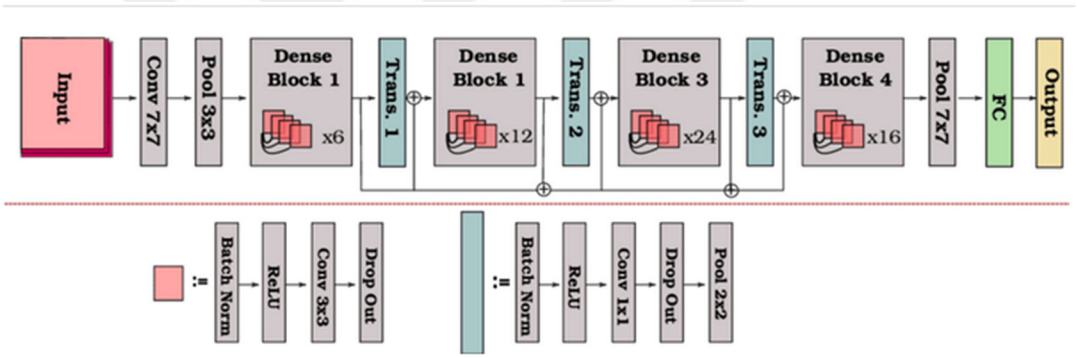


Figure 5.10: DenseNet121 Architecture.

As shown in the Figure 5.10 above, DenseNet121 is a deep network with four dense blocks, and the layer between two adjacent blocks is referred to as the transition layer. The transition layers are designed to change the size of the feature maps using convolutional neural networks and pooling operations.

5.3. Training and Validation

To assess the impact of the dataset on racial bias, several combinations of training and test sets are utilized, as outlined in Section 5.1. These combinations include two groups of datasets: those that are balanced (containing 2000 images per race) and those that are imbalanced (with 181,336 images for White, 21,103 for Asian, 19,582 for Black, and 14,227 for Indian). Each group has four datasets: Indian (I), Asian (A), White (W), and Black (B). To evaluate the effect of different combinations of datasets, all five models are

trained using 11 different combinations of the datasets, including W, B, A, I, IA, IB, IW, AB, AW, WB, and IAWB. Consequently, 11 x 5 = 55 training sessions are completed for each dataset group.

Implemented convolutional neural network-based models, including ResNet-50, InceptionV3, and DenseNet121, are pretrained in the source domain (AffectNet or ImageNet) in accordance with the specified requirements of the target domain to learn features.

The fine-tuning process for this study entails adjusting every layer of the model to the specific tasks of the target datasets. This technique functions as a regularizer that mitigates the effects of overfitting [104]. Throughout the training process, each layer is taught to perform the task classes within the target domain. Fine-tuning of this model utilized a learning rate of 0.001. The Adam optimizer is employed with batches set at 48, and the attenuation rate for fine-tuning training is set at 0.000001. If the accuracy of the validation set is not improved after three iterations, then the learning rate is reduced by 0.01. The evaluation criteria for this model employ the Top-1 Accuracy metric, which measures the proportion of accurately predicted samples. The evaluation is performed using 55 x 2=110 models trained with different combinations of balanced and imbalanced datasets. The train, validation, and test datasets are split in a ratio of 0.7, 0.2, and 0.1, respectively, for all trainings.

The experiments are executed on a personal computer with an Intel i9-12900H CPU, NVIDIA GeForce RTX 3080 Ti GPU, and 64.0 GB of RAM, the environment is Anaconda [105], and the operating system is Windows 10. Keras [106] and TensorFlow [107] are used to train the models.

6. RESULTS AND DISCUSSION

6.1. Balanced Dataset Experiments

The investigation into racial bias in the tested networks commenced with an assessment of existing bias. This is accomplished through training and evaluating each model 11 times using various combinations of datasets. To ensure an equal distribution of samples across each class, balanced datasets containing 2000 sample images per emotion are employed, as outlined in Section 5.1.1. Findings reveal that dataset imbalance does not significantly impact the results.

Deep-Emotion Network's results are demonstrated on Table 6.1. The first row indicates that when the model was trained on the White dataset, it performed the best on the White test set, but achieved inferior results on all other datasets during testing, suggesting a bias. This pattern was observed for all models trained on different datasets, implying that the method was unable to compensate for insufficient data from the missing races in the training set, resulting in lower accuracy. To eliminate racial bias, it is essential to train the model using a combination of datasets. For instance, training on both Indian and Asian images simultaneously enhanced the accuracy of all tests, including those of Indian and Asian. However, there is still a bias against the missing races. Finally, the bias is eliminated when training is conducted on all four datasets, as depicted in the final row. Therefore, the method is unbiased as long as the training data includes different races, and it does not compensate for the absence of samples from various races during training.

ACCURACY %		TEST										
		W	B	A	I	IA	IB	IW	AB	AW	WB	IAWB
TRAIN	W	64	34	39	36	37	36	53	36	54	51	45
	B	37	63	44	48	45	59	41	54	38	50	46
	A	41	48	65	50	57	48	47	57	56	41	48
	I	35	47	44	65	58	59	54	46	41	41	48
	IA	35	48	63	65	64	57	54	56	53	43	56
	IB	34	64	35	63	57	67	58	56	41	54	56
	IW	61	48	35	61	55	54	63	40	58	59	54
	AB	33	63	65	44	56	56	47	64	54	53	53
	AW	63	42	65	40	58	42	54	53	65	54	54
	WB	63	61	43	44	40	54	53	54	54	63	56
	IAWB	62	61	62	62	65	65	64	61	62	63	64

Table 6.1: Deep-Emotion results on 2000 images per emotion.

As shown in Table 6.2, the outcomes for the Self-Cure Network differ from those for Deep-Emotion. Specifically, the Self-Cure Network demonstrates a higher overall

accuracy and lower bias than Deep-Emotion. However, a slight bias is evident among the models trained and evaluated using a combination of Asian, Indian, and Black datasets. Minimizing bias is achieved by training the models on only two datasets: White and Black. The Self-Cure Network shows a more effective generalization than Deep-Emotion. It is important to note that bias still exists in the results, as the model trained on the White dataset shows a bias towards White test samples. The accuracy of White test samples decreased when the model is trained without White samples, resulting in a persistent bias towards the White dataset. However, the model trained on all datasets operates without bias, similar to Deep-Emotion.

ACCURACY %		TEST										
		W	B	A	I	IA	IB	IW	AB	AW	WB	IAWB
TRAIN	W	63	37	33	30	31	33	47	35	48	50	41
	B	47	60	55	57	56	58	52	58	51	48	55
	A	43	56	60	60	60	58	51	58	51	39	54
	I	36	53	47	66	56	60	51	50	41	39	50
	IA	39	60	59	64	62	63	52	60	49	41	56
	IB	43	62	59	64	62	63	53	61	51	52	57
	IW	60	55	57	65	61	60	63	56	58	56	59
	AB	46	53	63	61	63	61	53	62	55	51	58
	AW	62	40	61	54	58	53	58	57	62	52	58
	WB	60	59	53	58	56	58	59	57	57	60	58
	IAWB	63	63	63	66	65	64	65	63	63	60	64

Table 6.2: Self-Cure Network (SCN) results on 2000 images per emotion.

Adopted transfer learning methods: ResNet50 [80], DenseNet121 [81], and InceptionV3 [82] are followed a similar pattern to that of the Self-Cure Network, as demonstrated in Tables 6.3, 6.4, and 6.5. Similar to the Self-Cure Network, models trained solely on the White dataset exhibited a significant bias. However, when trained on combinations of the Asian, Indian, and Black datasets, the bias towards each other decreases.

ACCURACY %		TEST										
		W	B	A	I	IA	IB	IW	AB	AW	WB	IAWB
TRAIN	W	55	32	32	32	32	32	43	32	43	43	38
	B	38	53	49	51	51	52	45	51	43	46	48
	A	40	51	53	54	54	52	47	52	46	45	49
	I	36	50	43	59	51	55	47	47	40	43	47
	IA	37	48	52	55	54	52	46	51	44	43	48
	IB	39	53	49	55	52	54	47	51	44	46	49
	IW	52	47	47	53	50	50	53	47	49	49	50
	AB	39	50	44	51	47	50	44	47	42	44	46
	AW	53	45	48	48	48	47	51	47	51	49	49
	WB	49	43	35	40	37	41	44	47	42	46	41
	IAWB	48	46	51	53	51	49	51	48	49	47	49

Table 6.3: ResNet50 results on 2000 images per emotion.

ACCURACY %		TEST										
		W	B	A	I	IA	IB	IW	AB	AW	WB	IAWB
TRAIN	W	66	41	37	33	35	37	37	39	51	51	44
	B	38	46	44	46	45	46	46	45	42	42	43
	A	46	50	57	61	59	55	55	53	51	51	53
	I	37	56	48	66	56	61	61	51	42	42	51
	IA	42	61	62	64	63	62	62	61	51	51	57
	IB	40	54	60	51	55	61	61	56	50	50	51
	IW	63	55	51	64	57	59	59	53	57	57	58
	AB	42	61	53	64	61	62	62	59	50	50	56
	AW	63	53	50	58	56	56	56	53	58	58	57
	WB	66	57	61	51	51	54	54	53	57	57	56
	IAWB	59	60	57	59	60	59	59	60	60	60	60

Table 6.4: DenseNet121 results on 2000 images per emotion.

ACCURACY %		TEST										
		W	B	A	I	IA	IB	IW	AB	AW	WB	IAWB
TRAIN	W	59	37	42	38	40	37	49	39	51	48	44
	B	44	59	47	54	51	56	49	53	46	51	51
	A	40	55	55	60	57	57	50	55	47	47	52
	I	42	56	53	66	59	61	53	55	47	50	54
	IA	42	58	56	66	61	62	54	57	49	50	56
	IB	43	59	51	62	56	61	52	55	47	51	54
	IW	62	55	61	63	62	59	63	58	61	59	61
	AB	40	60	59	62	61	61	51	60	50	50	56
	AW	66	54	54	59	56	56	62	54	60	60	58
	WB	60	58	56	53	55	56	56	57	58	59	56
	IAWB	64	60	62	66	64	63	65	61	63	62	63

Table 6.5: InceptionV3 results on 2000 images per emotion.

6.2. Imbalanced Dataset Experiments

A further experiment is conducted to evaluate the effects of imbalanced datasets. In this experiment, the same training and testing procedures are followed; however, full datasets are utilized instead of balanced datasets (see Table 5.2). As stated in Section 5.1.1, the complete datasets comprise more samples. Although the distribution of samples among the classes is not balanced, the dataset suffers from bias in class accuracies. Furthermore, a larger sample count in the full datasets facilitated the learning process for the models.

As shown in Tables 6.6, 6.7, 6.8, 6.9, and 6.10 below, higher accuracy levels are revealed than those reported in the preceding experiment. However, a more pronounced bias is observed in this case. As the data count increases, the models can be optimized more effectively, enabling the weights to align with the data more accurately. Consequently, the model is better equipped to delineate the dataset in detail, owing to the substantial number of samples. The findings from the SCN and transfer learning methods indicate

that models trained on Indian, Asian, and Black datasets now exhibit a more discernible bias compared to the previous experiment.

ACCURACY %		TEST										
		W	B	A	I	IA	IB	IW	AB	AW	WB	IAWB
TRAIN	W	78	42	46	48	47	46	65	43	63	52	52
	B	31	77	51	62	56	68	48	59	41	50	57
	A	32	48	83	61	70	58	45	65	50	41	55
	I	31	54	56	84	76	65	48	50	36	42	58
	IA	31	61	81	82	84	64	63	65	63	49	65
	IB	32	74	59	81	61	79	62	64	47	63	64
	IW	77	53	55	81	63	65	81	46	62	63	63
	AB	32	72	80	63	63	63	48	81	62	64	66
	AW	77	42	79	54	64	48	64	65	80	61	64
	WB	74	72	50	55	45	65	64	65	62	82	65
	IAWB	74	75	77	77	77	76	76	76	75	75	77

Table 6.6: Deep-Emotion results on full datasets.

ACCURACY %		TEST										
		W	B	A	I	IA	IB	IW	AB	AW	WB	IAWB
TRAIN	W	81	41	46	49	47	45	66	43	64	62	55
	B	34	78	65	68	67	73	50	72	49	55	60
	A	33	70	91	73	83	72	52	80	60	50	66
	I	29	73	63	93	78	83	59	68	45	49	63
	IA	30	73	91	92	91	83	59	82	58	50	70
	IB	22	76	68	92	80	84	55	72	44	47	63
	IW	76	66	60	92	76	79	84	62	69	72	73
	AB	35	75	89	76	83	76	55	82	60	54	68
	AW	80	56	91	71	81	63	76	73	86	69	74
	WB	78	75	62	70	66	73	74	69	71	77	72
	IAWB	80	75	90	92	91	84	86	83	85	78	85

Table 6.7: Self-Cure Network (SCN) results on full datasets.

ACCURACY %		TEST										
		W	B	A	I	IA	IB	IW	AB	AW	WB	IAWB
TRAIN	W	80	27	43	46	45	36	64	34	63	55	50
	B	33	74	64	70	67	73	51	70	47	53	60
	A	37	51	88	69	78	60	52	70	60	44	60
	I	24	72	63	92	77	82	56	67	42	47	61
	IA	35	71	90	92	91	82	61	80	60	52	71
	IB	28	78	70	92	81	85	58	73	47	51	65
	IW	81	71	66	93	79	82	86	68	73	76	77
	AB	30	75	90	76	83	76	52	87	58	51	67
	AW	89	79	89	69	79	60	75	70	85	67	73
	WB	79	75	65	69	67	72	74	70	73	77	73
	IAWB	78	75	88	92	90	84	85	82	83	77	84

Table 6.8: Resnet50 results on full datasets.

ACCURACY %		TEST										
		W	B	A	I	IA	IB	IW	AB	AW	WB	IAWB
TRAIN	W	70	41	40	46	43	44	59	40	56	57	51
	B	21	74	63	68	65	71	43	69	40	46	55
	A	34	63	90	73	81	68	52	76	60	48	63
	I	26	73	67	91	79	82	56	70	45	48	63
	IA	32	73	89	91	90	82	60	81	59	52	70
	IB	23	75	70	90	80	83	54	73	44	48	63
	IW	78	71	69	91	80	81	85	70	75	75	77
	AB	31	73	88	74	81	74	52	81	58	51	65
	AW	79	58	89	71	80	64	75	74	84	69	75
	WB	64	69	58	62	60	65	63	63	62	66	63
	IAWB	75	74	86	90	88	82	83	80	81	75	81

Table 6.9: DenseNet121 results on full datasets.

ACCURACY %		TEST										
		W	B	A	I	IA	IB	IW	AB	AW	WB	IAWB
TRAIN	W	74	31	42	45	44	39	61	37	59	54	49
	B	24	75	61	66	63	70	44	67	40	48	55
	A	48	60	85	66	75	63	56	72	64	53	63
	I	22	73	66	92	80	82	55	71	44	46	63
	IA	34	73	88	92	90	82	61	80	59	52	70
	IB	30	76	64	89	77	83	58	71	46	52	64
	IW	79	68	65	91	78	80	85	67	73	74	76
	AB	32	75	86	74	80	75	52	86	57	52	65
	AW	79	55	88	69	79	63	75	72	84	68	74
	WB	77	75	62	69	65	72	74	68	70	76	71
	IAWB	75	72	86	90	88	81	82	79	80	74	81

Table 6.10: Inception v3 results on full datasets.

Consequently, Deep-Emotion exhibits a clear bias influenced by the data on which it has been trained on. The Self-Cure Network and transfer learning methods, particularly ResNet50, also exhibit bias towards the White dataset, although there are slight differences when using other datasets. Introducing variety by using only certain races during training helps, but it does not completely eliminate bias for all races not included in the training. To eliminate racial bias completely, all races must be represented equally during training. It is important to note that the tendency of racial bias in selected methods is unaffected by imbalanced classes, but increasing the dataset size can impact racial bias.

7. CONCLUSION

Due to the aforementioned considerations, an investigation of the influence of racial bias on facial expression recognition is conducted by compiling datasets from various racial backgrounds. These datasets are then utilized in different combinations to determine how the racial balance in the training data affects the model's racial bias. Analysis of the experimental results is conducted for samples from four distinct racial groups, namely White, Black, Indian, and Asian, both when combined and individually. To demonstrate the effectiveness of this approach, advanced facial expression recognition techniques and transfer learning methods, such as Deep Emotion, Self-Cure Network, ResNet50, InceptionV3, and DenseNet121, are employed as proof of concept.

The racial bias analysis in this study yielded three key findings. Initially, the examined methods exhibited a bias toward the races present in the training data. To eliminate racial bias, missing races are included in the training phase. Second, it is feasible to reduce racial bias by employing only a few races using specific techniques. In other words, even if certain races are not included in the training, the method can compensate for them and learn without bias, although this is not always possible. Finally, an improvement in performance does not necessarily indicate a reduction in racial bias. Conversely, the bias becomes more apparent as the networks adapt to the data. It is essential to note that the data utilized, and the selected methods are exclusive to this particular study.

In future work, this study can be extended by incorporating additional datasets and methodologies.

REFERENCES

[1] Roychowdhury, S., Emmons, M.: *A survey of the trends in facial and expression recognition databases and methods*. arXiv:1511.02407 (2015).

[2] Li, S., Deng, W.: *Deep facial expression recognition: a survey*. In: IEEE Transactions on Affective Computing. IEEE, pp 1–20 (2020).

[3] Kaminska, D., Aktas, K., Rizhinashvili, D., Kuklyanov, D., Sham, A.H., Escalera, S., Nasrollahi, K., Moeslund, T.B., Anbarjafari, G.: *Two-stage recognition and beyond for compound facial emotion recognition.* Electronics 10(22), 2847 (2021).

[4] Sang, D.V., Van Dat, N., et al.: *Facial expression recognition using deep convolutional neural networks*. In: 2017 9th International Conference on Knowledge and Systems Engineering (KSE), pp. 130–135. IEEE (2017).

[5] Drozdowski, P., Rathgeb, C., Dantcheva, A., Damer, N., Busch, C.: *Demographic bias in biometrics: a survey on an emerging challenge*. IEEE Trans. Technol. Soc. 1(2), 89–103 (2020).

[6] Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Menge- sha, Z., Toups, C., Rickford, J.R., Jurafsky, D., Goel, S.: *Racial disparities in automated speech recognition.* Proc. Natl. Acad. Sci. 117(14), 7684–7689 (2020).

[7] Xu, T., White, J., Kalkan, S., Gunes, H.: *Investigating bias and fairness in facial expression recognition.* In: European Conference on Computer Vision, pp. 506–523. Springer (2020).

[8] Yang, K., Qinami, K., Fei-Fei, L., Deng, J., Russakovsky, O.: *Towards fairer datasets: filtering and balancing the distribution of the people subtree in the imagenet hierarchy.* In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 547–558 (2020).

[9] De Vries, T., Misra, I., Wang, C., Van der Maaten, L.: *Does object recognition work for everyone?* In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Work- shops, pp. 52–59 (2019).

[10] Kasapoglu, T., Masso, A.: *Attaining security through algorithms: perspectives of refugees and data experts.* In: Theorizing Criminality and Policing in the Digital Media Age. Emerald Publishing Limited (2021).

[11] Perkowitz, S.: *The bias in the machine: Facial recognition technology and racial disparities.* MIT Case Studies in Social and Ethical Responsibilities of Computing, no. Winter. https://mit-serc.pubpub.org/pub/bias-in-machine (2021).

[12] Robinson, J.P., Livitz, G., Henon, Y., Qin, C., Fu, Y., Timoner, S.: *Face recognition: too bias, or not too bias?* In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–1 (2020).

[13] Das, A., Dantcheva, A., Bremond, F.: *Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach.* In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops (2018).

[14] Guo, G., Mu, G.: *Human age estimation: What is the influence across race and gender?* In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pp. 71–78. IEEE (2010).

[15] Chen, Y., Joo, J.: *Understanding and mitigating annotation bias in facial expression recognition.* In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 14980– 14991 (2021).

[16] Li, S., Deng, W., Du, J.: *Reliable crowdsourcing and deep locality- preserving learning for expression recognition in the wild.* In: Proceedings of the IEEE Conference on Computer Vision and Pat- tern Recognition, pp. 2852–2861 (2017).

[17] Szegedy C., Ioffe S., Vanhouche V., and Alemi A. *Inception-v4, inception-resnet and the impact of residual connections on learning.* Arxiv, 1602.07261, (2016).

[18] Hinton G.E., Srivastava N., Krizhevsky A., Sutskever I., and Salakhutdinov R. *Improving neural networks by preventing co-adaptation of feature detectors.* CoRR, abs/1207.0580, (2012).

[19] T. Hastie, R. Tibshirani, and J Friedman. *The elements of statistical learning: data mining, inference, and prediction.* Springer, Second edition, (2009).

[20] Washington, P., Kalantarian, H., Kent, J., Husic, A., Kline, A., Leblanc, E., Hou, C., Mutlu, C., Dunlap, K., Penev, Y., et al. *Training affective computer vision models by crowdsourcing soft-target labels*. Cognitive computation, 13:1363–1373, (2021).

[21] Elman J.L. *Finding structure in time*. Cognitive science, 14(2):179–211, (1990).

[22] D. Kingma and J. Ba. Adam: *a method for stochastic optimization* Arxiv, 1412.6980, (2014).

[23] A. Krizhevsky, I. Sutskever, and G.E. Hinton. *Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems,* pages 1097–1105, (2012).

[24] Jordan M.I. *Artificial Neural Network*, pages 112-127. IEEE Press, (1990).

[25] Y. Nesterov. *A method of solving a complex programming problem with convergence rate o(1/k2)*. Soviet Mathematics Doklady, 27:372– 376, (1983).

[26] B.T. Polyak. *Some methods of speeding up the convergence of iteration methods.* USSR Computational Mathematics and Mathematical Physics, 4(5):1–17, (1964).

[27] Hochreiter S. and SchmidhuberJ. *Long short-term memory. Neural Computation*, 9(8):1735–1780, (1997).

[28] I. Sutskever, J. Martens, G.E. Dahl, and G.E. Hinton. *On the importance of initialization and momentum in deep learning*. ICML, 28(3):1139– 1147, (2013).

[29] LeCun Y., Bottou L., Bengio Y., and Haffner P. *Gradient-based learning applied to document recognition*. IEEE Communications magazine, 27(11):41–46, (1998).

[30] LeCun Y., Jackel L., Boser B., Denker J., Graf H., Guyon I., Henderson D., Howard R., and Hubbard W. Handwritten digit recognition: *Applications of neural networks chips and automatic learning*. Proceedings of the IEEE, 86(11):2278–2324, (1998).

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

[32] Ali, G., Iqbal, M. A., and Choi, T.-S. *Boosted nne collections for multicultural facial expression recognition*. Pattern Recognition, 55:14–27, (2016).

[33] Barocas, S., Hardt, M., and Narayanan, *A. Fairness and Machine Learning: Limitations and Opportunities.* fairml-book.org, (2019). http://www.fairmlbook.org.

[34] Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., and et al., S. H. *AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias.* CoRR, abs/1810.01943, 2018. URL http://arxiv.org/abs/1810.01943.

[35] Bird, S., Dudik, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., and Walker, K. Fairlearn: *A toolkit for assessing and improving fairness in AI.* Technical Report MSR-TR-2020-32, Microsoft, May (2020).

[36] Bradski, G. The OpenCV Library. Dr. Dobb's Journal of Software Tools, 2000. Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., and Cosentini, A. C. *The zoo of fairness metrics in machine learning*. 2021.

[37] Chen, Y. and Joo, J. Understanding and mitigating annotation bias in facial expression recognition. (arXiv:2108.08504), Aug 2021a. URL:http://arxiv. org/abs/2108.08504. Aaron Smith and Janna Anderson. "AI, Robotics, and the Future of Jobs". In: Pew Research Center 6 (2014), p. 51.

[38] Emanuele Neri et al. *Artificial intelligence: Who is responsible for the diagnosis?* (2020).

[39] Jean-Francois Bonnefon, Azim Shariff, and Iyad Rah- wan. *"The social dilemma of autonomous vehicles"*. In: Science 352.6293 (2016), pp. 1573–1576.

[40] James Zou and Londa Schiebinger. *AI can be sexist and racist—it's time to make it fair.* (2018).

[41] Eric Mack. Hawking, Musk, Wozniak *Warn About Artificial Intelligence's Trigger Finger*. https://www.forbes.com/sites/ericmack/2015/07/27/hawking-musk-wozniak-freaked-about-artificial-intelligence-getting-a-trigger-finger/?sh=7ad6f69f7416.

[42] Frank Rosenblatt. *"The perceptron: a probabilistic model for information storage and organization in the brain.*" In: Psychological review 65.6 (1958), p. 386.

[43] Seppo Linnainmaa. *"The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors"*. In: Master's Thesis (in Finnish), Univ. Helsinki (1970), pp. 6–7.

[44] Ning Qian and Terrence J Sejnowski. *"Predicting the secondary structure of globular proteins using neural network models"*. In: Journal of molecular biology 202.4 (1988), pp. 865–884.

[45] algorithmwatch.org. *Finnish Credit Score Ruling raises Questions about Discrimination and how to avoid it.* https:/ algorithmwatch.org/en/story/finnish-credit-score-ruling-raises-questions-about-discrimination-and-how-to-avoid-it/.

[46] Daisuke Wakabayashi. *Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam.* https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html.

[47] Niraj Chokshi. *Tesla Autopilot System Found Probably at Fault in 2018 Crash*. https://www.nytimes.com/ 2020/02/25/business/tesla-autopilot-ntsb.html.

[48] Julia Angwin. *Facebook Enabled Advertisers to Reach 'Jew Haters'*. https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters.

[49] C. Darwin. *The Expression of Emotions in Man and Animals*. John Murray, reprinted by University of Chicago Press, (1965).

[50] P. Ekman. *The Argument and Evidence about Universals in Facial Expressions of Emotion*, pages 143–164. Wiley, New York, (1989).

[51] P. Ekman and W. Friesen. The Facial Action Coding System*: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, San Francisco, (1978).

[52] J. Carroll and J. Russell. *Facial expression in hollywood's portrayal of emotion. Journal of Personality and Social Psychology*., 72:164–176, (1997).

[53] C. Izard, L. Dougherty, and E. A. Hembree. *A system for identifying affect expressions by holistic judgments*. In Unpublished Manuscript, University of Delaware, (1983).

[54] B. Fasel and J. Luttin. *Recognition of asymmetric facial action unit activities and intensities*. In Proceedings of International Conference of Pattern Recognition, (2000).

[55] Eihl-Eihesfeldt. Human Ethology. Aldine de Gruvter, New York, 1989. T. Kanade, J. Cohn, and Y.-L. Tian. *Comprehensive database for facial expression analysis*. In Proceedings of International Conference on Face and Gesture Recognition, pages 46–53, (2000).

[56] W. Friesen and P. Ekman. *Emfacs-7: emotional facial action coding system. Unpublished manuscript, University of California at San Francisco*, (1983).

[57] R. Matias, J. Cohn, and S. Ross. *A comparison of two systems to code infants' affective expression.* Developmental Psychology, 25:483–489, (1989).

[58] Liu, Z., Luo, P., Wang, X., Tang, X.: *Deep learning face attributes in the wild*. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3730–3738 (2015).

[59] Fu, S., He, H., Hou, Z.-G.: *Learning race from face: a survey*. IEEE Trans. Pattern Anal. Mach. Intell. 36(12), 2483–2509 (2014).

[60] Domnich, A., Anbarjafari, G.: Responsible ai: *gender bias assessment in emotion recognition*. arXiv:2103.11436 (2021).

[61] Conley, M.I., Dellarco, D.V., Rubien-Thomas, E., Cohen, A.O., Cervera, A., Tottenham, N., Casey, B.: *The racially diverse affective expression (radiate) face stimulus set*. Psychiatry Res. 270, 1059– 1067 (2018).

[62] Dailey, M.N., Joyce, C., Lyons, M.J., Kamachi, M., Ishi, H., Gyoba, J., Cottrell, G.W.: *Evidence and a computational explanation of cultural differences in facial expression recognition*. Emotion 10(6), 874 (2010).

[63] Fischer, A.H., Rodriguez Mosquera, P.M., Van Vianen, A.E., Manstead, A.S.: *Gender and culture differences in emotion*. Emotion 4(1), 87 (2004).

[64] Laurence, S., Zhou, X., Mondloch, C.J.: *The flip side of the other- race coin: they all look different to me*. Br. J. Psychol. 107(2), 374–388 (2016).

[65] Prado, C., Mellor, D., Byrne, L.K., Wilson, C., Xu, X., Liu, H.: *Facial emotion recognition: a cross-cultural comparison of Chinese, Chinese living in Australia, and Anglo-Australians*. Motiv. Emot. 38(3), 420–428 (2014).

[66] Strohminger, N., Gray, K., Chituc, V., Heffner, J., Schein, C., Hea- gins, T.B.: *The mr2: a multi-racial, mega-resolution database of facial stimuli.* Behav. Res. Methods 48(3), 1197–1204 (2016).

[67] Shimoda, K., Argyle, M., Bitti, P.R.: *The intercultural recognition of emotional expressions by three national racial groups: English, Italian and Japanese*. Eur. J. Soc. Psychol. 8(2), 169–179 (1978).

[68] Ghallab, M.: *Responsible ai: requirements and challenges*. AI Perspect. 1(1), 1–7 (2019).

[69] Benjamins, R., Barbado, A., Sierra, D.: *Responsible ai by design in practice*. arXiv:1909.12838 (2019).

[70] Vetrò, A., Santangelo, A., Beretta, E., De Martin, J.C.: *Ai: from rational agents to socially responsible agents.* Digital Policy, Regulation and Governance (2019).

[71] Livingston, M.: *Preventing racial bias in federal ai*, JSPG, vol. 16 (2020)

[72] Shneiderman, B.: *Responsible ai: bridging from ethics to practice*. Commun. ACM 64(8), 32–35 (2021).

[73] Wang, W., He, F., Zhao, Q.: *Facial ethnicity classification with deep convolutional neural networks*. In: Chinese Conference on Biometric Recognition, pp. 176–185. Springer (2016).

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

[75] Benitez-Garcia, G., Nakamura, T., Kaneko, M.: *Multicultural facial expression recognition based on differences of Western-Caucasian and East-Asian facial expressions of emotions.* IEICE Trans. Inf. Syst. 101(5), 1317–1324 (2018).

[76] Sohail, M., Ali, G., Rashid, J., Ahmad, I., Almotiri, S.H., AlGhamdi, M.A., Nagra, A.A., Masood, K.: *Racial identity-aware facial expression recognition using deep convolutional neural networks*. Appl. Sci. 12(1), 88 (2022).

[77] Olteanu, A., Castillo, C., Diaz, F., Kıcıman, E.: *Social data: biases, methodological pitfalls, and ethical boundaries.* Front. Big Data 2, 13 (2019).

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

[79] Kai, W., Xiaojiang, P., Jianfei, Y., Shijian, L., Yu, Q: *Suppressing uncertainties for large-scale facial expression recognition*. arXiv:2002.10392 (2020).

[80] He, K., Zhang, X., Ren, S., Sun, J.: *Deep residual learning for image recognition.* (2015).

[81] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: *Rethinking the inception architecture for computer vision* (2015).

[82] Huang, G., Liu, Z., Weinberger, K.Q.: *Densely connected convolutional networks.* CoRR, vol. abs/1608.06993. http://arxiv.org/abs/ 1608.06993 (2016).

[83] Hochreiter, S.: *The vanishing gradient problem during learning recurrent neural nets and problem solutions*. Int. J. Uncertain. Fuzzin. Knowl. Based Syst. 6(02), 107–116 (1998)

[84] Buolamwini, J., & Gebru, T. Gender Shades: *Intersectional Accuracy Disparities in Commercial Gender Classification.* Proceedings of Machine Learning Research, 81, 77-91, (2018).

[85] Ayanna Howard, Cha Zhang, and Eric Horvitz. *"Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems"*. In: 2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO). IEEE. 2017, pp. 1–7

[86] Gera, D., & Balasubramanian, S. *Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition*. Pattern Recognition Letters, 145, 58-66, (2021).

[87] ZHAN, Fangneng, et al. *Multimodal image synthesis and editing: A survey and taxonomy*. IEEE Transactions on Pattern Analysis and Machine Intelligence, (2023).

[88] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. *A Survey on Bias and Fairness in Machine Learning.* ACM Comput. Surv. 54, 6, Article 115, 35 pages, (2022).

[89] Oneto, Luca; CHIAPPA, Silvia. *Fairness in machine learning*. In: Recent trends in learning from data: Tutorials from the inns big data and deep learning conference (innsbddl2019). Springer International Publishing. p. 155-196, (2020).

[90] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. *A survey on bias and fairness in machine learning*. ACM computing surveys (CSUR), 54(6), 1-35, (2021).

[91] Rhue, Lauren, *Racial Influence on Automated Perceptions of Emotions* (November 9, 2018). Available at SSRN: https://ssrn.com/abstract=3281765

[92] Bellamy, Rachel & Dey, Kuntal & Hind, Michael & Hoffman, Samuel & Houde, Stephanie & Kannan, Kalapriya & Lohia, Pranay & Martino, Jacquelyn & Mehta, Sameep & Mojsilovic, Aleksandra & Nagar, Seema & Natesan Ramamurthy, Karthikeyan & Richards, John & Saha, Diptikalyan & Sattigeri, Prasanna & Singh, Moninder & Varshney, Kush & Zhang, Yunfeng. *AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias.* IBM Journal of Research and Development. PP. 1-1. 10.1147/JRD.2019.2942287, (2019).

[93] Peng, Z., Li, J., and Sun, Z. *Emotion recognition using generative adversarial networks.* In 2020 International Conference on Computer Engineering and Intelligent Control (ICCEIC), pp. 77–80, (2020). doi: 10.1109/ICCEIC51584. 2020.00023.

[94] Porcu, S., Floris, A., and Atzori, L. *Evaluation of data augmentation techniques for facial expression recognition systems*. Electronics, 9(11), (2020). ISSN 2079-9292. doi: 10.3390/electronics9111892. URL https://www. mdpi.com/2079-9292/9/11/1892

[95] Khorrami, P., Paine, T., and Huang, T. *Do deep neural networks learn facial action units when doing expression recognition?* In Proceedings of the IEEE international conference on computer vision workshops, pp. 19–27, (2015).

[96] Ko, B. *A brief review of facial emotion recognition based on visual information*. Sensors (Basel, Switzerland), 18, (2018).

[97] Mollahosseini, A., Hasani, B., and Mahoor, M. H. *Affectnet: A database for facial expression, valence, and arousal computing in the wild*. IEEE Transactions on Affective Computing, (2017).

[98] LoBue, V. and Thrasher, C. *The child affective facial expression (cafe) set: validity and reliability from untrained adults*. Front. Psychol., 5, (2015).

[99] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. *The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotionspecified expression.* In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, pp. 94–101, (2010).

[100] Lyons, M., Akamatsu, S., Kamachi, M., and Gyoba, J. *Coding facial expressions with gabor wavelets*. In Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200–205, (1998).

[101] Zhu, X., Li, L., Zhang, W., Rao, T., Xu, M., Huang, Q., and Xu, D. Dependency exploitation: A unified cnn-rnn approach for visual emotion recognition. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, pp. 3595–3601, (2017).

[102] Hasani, B. and Mahoor, M. H. *Facial expression recognition using enhanced deep 3d convolutional neural networks*. In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, Jul (2017).

[103] Karkkainen, K. and Joo, J. Fairface: *Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation*. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1548–1558, (2021).

[104] Sham, A.H., Aktas, K., Rizhinashvili, D. et al. *Ethical AI in facial expression analysis: racial bias*. SIViP 17, 399–406 (2023).

[105] Anon, 2020. Anaconda Software Distribution, Anaconda Inc. Available at: https://docs.anaconda.com/.

[106] Chollet, F. & others, 2015. Keras. Available at: https://github.com/fchollet/keras.

[107] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Murray, Chris Olah, Jonathon Shlens, Benoit Steiner, Ilya Sutskever,. *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015. Software available from tensorflow.org.