Republic of Iraq

Ministry of Higher Education and Science Research

University of Babylon

College of Science for Women

Department of Computer Science

Recognizing Emotional Facial Expression of Autism Based on Deep Learning

A Thesis

Submitted to the council of College of Science for woman,

University of Babylon in Partial Fulfillment of the Requirement

For Degree of Master of Science in Computer Science

By

Mays Ali Shaker

Supervised by

Asst. Prof. Dr. Amina Atiya Dawood

2023 A.D. **1445 A.H.**

فَتَعَٰلَى ٱللَّهُ ٱلۡمَلِكُ ٱلۡحَقُّۗ وَلَا تَعۡجَلۡ بِٱلۡقُرۡءَانِ مِن قَبۡلِ
أَن يُقۡضَىٰٓ إِلَيۡكَ وَحۡيُهُۥۖ وَقُل رَّبِّ زِدۡنِي عِلۡمٗا ١١٤

صدَقَ الله العَظِيم

سُورَةُ طه

CERTIFICATION OF THE EXAMINATION COMMITTEE

We are the members of the examination committee, certify that we have read this thesis entitled (**Recognizing Emotional Facial Expression of Autism Based on Deep Learning**) and after examining the master student (**Mays Ali Shaker**) in its contents 12/10/2023 and that in our opinion it is adequate as a thesis for the degree of Master in Science /Computer Science with degree (Excellence).

Committee Chairman

Signature:
Name: **Muhammed Abaid Mahdi**
Scientific Order: **Assist. Prof. Dr.**
Address: University of Babylon / College of Science for Women
Date: / /2023

Committee Member

Signature:
Name: **Asaad Noori Hashim**
Scientific Order: **Assist. Prof. Dr.**
Address: University of Kufa / College of Computer Science and Mathematics
Date: / / 2023

Committee Member

Signature:
Name: **Farah Mohammed Hassan Al-Shareefi**
Scientific Order : **Dr.**
Address: University of Babylon/ College of Science for Women
Date: / / 2023

Committee Member (Supervisor)

Signature:
Name: **Amina Atiya Dawood**
Scientific Order: **Assist. Prof. Dr.**
Address: University of Babylon/ Presidency of the University of Babylon / Computer Center
Date: / / 2023

Date of Examination: 12/10/2023
Deanship Authentication of College of Science for Woman
Approved for the College Committee of graduate studies.

Signature:
Name: **Abeer Fauzi Al-Rubaye**
Scientific Order : **Prof. Dr.**
Address: **Dean of College of Science for Woman**
Date: / / 2023

Supervisors Certification

*We certify that this thesis entitled "**Recognizing Emotional Facial Expression of Autism Based on Deep Learning**", is completed by the student "Mays Ali Shaker" under my supervision at the Department of Computer Science of University of Babylon in a partial fulfillment of the requirements for MSc Degree in Computer Science.*

Signature:

Name: Asst. Prof. Dr.Amina Atiya Dawood

Date: / / 2023

Address: University of Babylon /Presidency of University of Babylon / Computer Center

The Head of the Department Certification

*In view of the available recommendations, I forward the dissertation entitled "**Recognizing Emotional Facial Expression of Autism Based on Deep Learning**" for examining committee.*

Signature:

Name:. Asst. Prof. Dr. Saif Mahmood Khalaf

Date: / / 2023

Address: University of Babylon/College of Science for Women

Acknowledgments

All thanks and praise to Allah, the Lord of the world, who gave me courage and enabled me to achieve this work.

My thanks and gratitude to my supervisor ***Dr. Amina Atiya Dawood*** *for the support and guidance she have given me and the effort and time to complete this research.*

Thanks, and gratitude to all my professors and all the staff of the Department of Computer Sciences \ College of Sciences for Women \University of Babylon for their help.

All Thanks and Gratitude to all my family and friends for their support and encouragement.

Mays Ali ... (2023)

Dedication

To my great creator

*To the first teacher of mankind**, the prophet Mohammed** ,*

*To **my beloved father** who is my support in my life.*

*To **my dear mother** who has a kind heart.*

*To those who supported me and encouraged me with all love and patience.. **my dear aunt.***

*To my beloved **sister** and **brothers** who are always beside me.*

*To **my respected professors**.*

*To **everyone** who stood with me.*

Mays

Abstract

Emotion on our face can determine our feelings, mental state and can directly impact our decisions. Humans are subjected to undergo an emotional change in relation to their living environment and or at a present circumstance. These emotions can be disgust, fear, sadness, joy or anticipation. Due to the intricacy and nuance of facial expressions and their relationship to emotions, accurate facial expression identification remains a difficult undertaking. People with autism are characterized by atypical facial expressions, so they have difficulty communicating and interacting with others. Due to the significant increase in their numbers in recent years, researchers and scientists have been interested in developing tools that help them express their emotional state. but Unfortunately, most studies have focused on analyzing autistic behaviors using invasive tools. These tools trigger an unwanted response due to their sensitivity and need to control the environment under specific conditions. There is also a lack of works that relies on natural, spontaneous data to reveal the behaviors of people with autism.

Therefore, with developing in deep learning techniques, this thesis has developed an automated model to detect emotions for people with autism in real-time using Dynamic videos. The model was based on the importance of face, head, and eye gaze' role in interacting with other to be as an input for this model. The novel contribution of this work is investigating the differences of facial expression between people with autism and those without autism, during express the same emotions.

Using Googlenet algorithm contributed to produce significant results in feature extraction and emotion prediction. where the accuracy for model training and validation were 99.88% and 98.54% respectively, also The model proved its efficiency towards unseen data with an accuracy with 97.54%.

The Content

List of Figures

List of Tables

List of Abbreviations

Abbreviation	***Description***
ASD	*Autism Spectrum Disorder*
TD	*Typical Development*
AS	*Asperger Syndrome*
PDD-NOS	*Pervasive Developmental Disorder- Not Otherwise Specified*
AU	*Action Unit*
ACC	*Accuracy*
CSV	*Comma Separated Values*
DL	*Deep Learning*
HCL	*Human Computer Interaction*
PDF	*Probability Density Function*
AI	*Artificial Intelligence*
RELU	*Rectified Linear Units*
ANN	*Artificial Neural Network*
CM	*Confusion Matrix*
CNN	*Convolution Neural Network*
FC	*Fully-Connection*
FN	*False Negative*
FP	*False Positive*
LSTM	*Long Short-Term Memory*
ML	*Machine Learning*
RNN	*Recurrent Neural Network*
TN	*True Negative*
TP	*True Positive*

Chapter One

Introduction

1.1 Introduction

Facial expressions are powerful non-verbal cues that convey emotions and intentions. They can be categorized into six primary types: joy, sadness, fear, disgust, and anticipation. These expressions are universal, transcending language barriers[1].

However, individuals with autism often struggle with interpreting and displaying facial expressions. They may have difficulty recognizing subtle emotional cues, hindering their social interactions. People with autism might exhibit atypical expressions themselves, such as limited eye contact or head movement, making it challenging for others to gauge their emotions[2].

Autism Spectrum Disorder (ASD) is a group of neurodevelopmental disorders[3] characterized by difficulties with communication, social interaction, and repetitive behaviors[4]. Autism, Asperger Syndrome(AS), and Non-Specified Pervasive Developmental Disorder (PDD-NOS) are the three different categories of autism spectrum disorders. Nevertheless, these categories are presently referred to as (ASD) [5]. In this study, we will specifically concentrate on individuals with AS, which typically exhibit average or above-average intelligence and normal language development[6].There are currently no known interventions or treatments that can cure or fully alleviate (ASD) [7]. However, comprehensive behavioral and educational intervention programs were developed and play an important role in helping persons with ASD to overcome impairments in communication and social interaction[8].

One of the central topics in psychology and autism research is empathy[9]. Empathy is frequently defined as the capacity to comprehend the mental states of others, such as their emotions, thoughts, and intentions, and to respond with the appropriate emotions or expressions[10]. When we demonstrate empathy, we make response in ways that acknowledge the emotions and beliefs of others and are

sensitive to their points of view. Where understanding and managing emotions are vital for successful social engagement[11].

Technology has an increasing potential to bridge the gap between what individuals with autism feel internally and what they express externally. However, it is necessary to consider how to design this technology in a way that respects the human need to control the display of emotions. In 1997, the concept of affective Computing was introduced by Rosalind Picard. Affective computing is the study and development of computer systems and devices capable of recognizing, interpreting, processing, and simulating human emotions[12].

All of these methods, however, have focused on identifying and interpreting the facial expressions of individuals with autism spectrum disorder (ASD) through the use of wearable devices and physiological signals[13][14].But, the use of these technologies is restricted, and must be applied in a controlled environment[15].

In recent years, deep learning has gained significant popularity due to its remarkable performance in tasks related to problem classification and recognition. It can handle complex and high-dimensional data and can learn directly from raw data without relying on handcrafted features or explicit feature engineering[16][17].

This study aimes to developing an emotional model specifically for individuals with Asperger's syndrome (AS) by detecting and classifying their emotional stat es using facial expressions, eye gaze, and head gestures. The model used a modified GoogLeNet, due to its efficiency and feature learning capabilities [18][19].

The emotional model was able to generate predictions for time series data related to facial expressions, eye fixation, and head gestures of individuals with Asperger's syndrome (AS) by utilizing the modified GoogLeNet structure. This allowed the model to effectively analyze and interpret the emotional cues exhibited by these individuals over time. The predictions generated by the model provided valuable insights into the emotional states of individuals with AS, aiding in understanding their emotional expressions and potentially assisting in developing appropriate interventions or support strategies.

1.2 Related Works

This section will present the different works which have been developed by other researchers in emotions recognition area. The presenting works based on different aspects like the type of emotions extraction, differences in emotions, technical methods which used in emotions recognition, and input channels to extract emotions. For example

- (**Drimalla et al**., 2021) In their research, facial expressiveness in an imitation context was measured using a regression model. Where qualitative differences in facial expressions were measured between neurotypical and autistic individuals. A webcam was installed to monitor their level of concentration. Researchers assessed the participants' ability to replicate facial expressions they were instructed to imitate and to recognize emotions. Although autistic participants were able to imitate the facial expressions as instructed, their imitation was delayed and less accurate than that of neurotypical participants[20]. This study has limitations in that imitation does not appear to aid in real-time emotion recognition.
- (**Alqahtani et al**., 2021) , proposed integrating machine learning (SVM, DT) and physiological signals to determine the sentiments of students during a test. They collected EEG, ECG, and EMG data from 27 individuals taking a computerized English language test using wearable sensors[15]. Due to their invasiveness and inconvenient nature, physiological signal sensors may not be practicable for ITS users outside of a laboratory context.
- (**Silva et al**., 2021) they developed a system for automatic emotion detection through facial expressions analysis. The system, designed to interact with children diagnosed with Autism Spectrum Disorder (ASD) through the Zeno R50 Robokind® (ZECA) robot, utilized the Intel® RealSense™ 3D sensor for real-time facial expression recognition. By extracting facial features and using a multiclass Support Vector Machine classifier, the system achieved a 93.6%

accuracy rate[21]. However, there were limitations, particularly in accurately identifying 'Anger' and 'Fear' emotions during the IMITATE activity, possibly due to the children's difficulty in interpreting ZECA's(robot) facial expressions.

- (**Hassouneh et a**l., 2020), they developed a classifier that utilized facial expression feature points and electroencephalograph (EEG) signals. The classification of basic emotions was achieved using convolutional neural network (CNN) and long short-term memory (LSTM) algorithms. The study involved 35 males and 25 females as participants. Haar algorithm was employed to extract the feature points from the participants' faces. The accuracy rate achieved was 99.81% using CNN to detect emotions using facial landmarks and 87.25% to EEG signals using LSTM[22]. .However, the study had certain limitations including the lack of a sufficient number of available topics to collect data and the need for additional features from EEG signals.
- (**Singh and Dewan.,** 2020), developed a real-time system for detecting emotions in autistic children, which consisted of three main stages: face recognition, facial feature extraction, and feature classification. The detection algorithm(Viola-Jones) and Adaboost were utilized in this process. The system demonstrated the ability to identify seven different facial emotions, including anger, disgust, fear, joy, sadness, contempt, and surprise. Testing the system with a group of children aged between 6 and 14 years resulted positive outcomes, with an accuracy rate of 85.97%[23]. However, an inherent limitation of this study was that the algorithm provided a binary output, only indicating whether a face was recognized or not. Consequently, statistical validation of the experimental results was not possible.

- (**Manfredonia et al**., 2019) In their research, they focused on the ability of individuals with Autism Spectrum Disorder (ASD) to generate facial expressions of emotions when given verbal instructions. They employed the Janssen Autism

Knowledge Engine (JAKE®), which utilized automated facial expression analysis software (FACET) to assess facial expressions in two groups: individuals with ASD (n=144) and a typically developing (TD) comparison group (n=41). The study revealed disparities in the ability to produce facial expressions between the ASD and TD groups, specifically in terms of activating facial action units (AUs) associated with happiness, fear, surprise, and disgust, while AUs linked to anger or sadness did not show significant differences. The findings indicated that individuals with ASD demonstrated statistically significant variations in their capability to exhibit appropriate facial expressions when prompted to display certain emotions, as evidenced by reduced evidence across specific AUs. Notably, this was observed in AUs related to happiness (AU12), fear (AU5), surprise (AU5), and disgust (AU9), but not in any of the predetermined AUs associated with anger or sadness[24].

- (**Król & Król**., 2019) , investigated the variations in eye movement patterns between individuals with Autism Spectrum Disorder (ASD) and typically developing individuals (TD). The researchers recorded and compared the eye movements of 21 participants with autism and 23 participants without autism who were similar in terms of age, gender, and IQ. The main objective was to assess the importance of visual information obtained from facial stimuli in the detection of emotions. The study employed pictures depicting six different emotions (neutral, sad, disgusted, fearful, and happy). The researchers utilized the SMI event detector high-speed detection algorithm with default settings to analyze the participants' eye fixations. The findings revealed that individuals with autism tended to focus more on the lower facial regions and less on the eyes compared to the typically developing group. This suggests that a portion of the difficulty experienced by autistic individuals in identifying emotions can be attributed to the initial stage of face processing, specifically the extraction of visual information through eye fixations [25].

- (**Samad et al**., 2019), proposed a new approach to examine distinctive characteristics associated with Autism Spectrum Disorder (ASD) in spontaneous facial expressions, which can be distinguished from facial expressions of typically developing controls (TD) based on observable features. Advanced computer vision techniques were utilized to automatically monitor subtle facial movements, enabling quantitative analysis of behavioral abnormalities related to ASD. In order to collect facial data, audio and visual stimuli were employed, including the use of a computer-generated avatar and a PrimeSense sensor. The researchers utilized a commercially available real-time facial motion capture system called The faceshift studio, which was selected to encompass six common emotional expressions (happiness, sadness, fear, anger, surprise, disgust)[26]. Limitations of this study, is the use of a sensor that requires a controlled environment.

- (**Dawood et al**., 2018), developed an affective model employing deep learning algorithms (CNN, LSTM) to infer the cognitive-affective states of Asperger's syndrome pupils in real-time. Students interacted with a computer in an unconstrained environment while a natural-spontaneous affective dataset was collected from their facial expressions, head movement, and eye fixation. The model effectively detected five affective states (confidence, uncertainty, engagement, anxiety, and boredom) with an accuracy rate of 90.06 % [5].

Table (1.1): Summary of the Related Works

Author Name	Type of dataset	Tools that used to capture features	Methodology	Input channels	Acc.
Drimalla et al., 2021[20]	Induced Data	Webcam	regression model	Facial movements	-
Alqahtani et al., 2021[15]	Induced Data	wearable physiological sensors	machine learning(SVM, DT)	physiological signals (EMG EEG,ECG)	81%
Silva et al., 2021 [21]	Induced Data	1.camera 2. robot 3. sensor	Support Vector Machine classifier	-Action unit -Head movement	93.6%
Hassouneh et al., 2020[22]	Induced Data	1.webcam 2. EPOCheadset	-CNN -LSTM	- Action unit -EEG signals	-99.81% -87.25%
Singh and Dewan 2020[23]	Spontaneous Data	Camera	-Viola jones algorithm -adaboost	-AUs	85.97%
Manfredonia et al. 2019[24]	Induced Data	Camera	automated facial expression analysis software (FACET)	Action unit	-
Król & Król, 2019[25]	Induced Data	Camera(remote eye-tracking device SMI RED250Mobile)	the SMI event detector high-speed detection algorithm	Eye movement	-
Samad et al., 2019[26]	Induced Data	- PrimeSense sensor - The faceshift studio is a facial motion tracking	computer vision techniques	Action unit	-

Dawood et al.,2018[5]	Spontaneous Data	Webcam	CNN LSTM	-facial expressions -head movement -eye gaze	90.06%

1.3 Thesis Motivations

Based on studies which stated that people with AS prefer interact to with their colleagues through mobile or computer rather than interact with them face to face. Also, they spent almost their time interacting with computer or mobile. And due to the lack in tools that can captures their interaction without disturb them either with caregiver, friends or their families. The field of affective computing which enabled machines to interact with users, and developing deep learning techniques to serve as application in real world, motivated me to build this model. So, integrating computing field with psychological field was the seed for the work in this thesis.

1.4 Statement of The Problem

1- Previous research has focused on extracting emotions from individuals with autism using physiological data collected through wearable instruments.

However, these tools have limitations, including impracticality for real-time applications and invasiveness for individuals with autism.

2- Some studies have relied on manually interpreting induced emotions by human judges, which may not fully capture the dynamic nature of emotions and can lead to misclassification.

3- Other studies have primarily used static images displaying intense facial expressions, which may not accurately represent the full range of emotions expressed by individuals with ASD.

Therefore, there is a need to develop a model that utilizes dynamic video stimuli to capture a wider range of emotions expressed by individuals with ASD without requiring user intervention, and it should be applicable in real-time scenarios.

1.5 The Aim of Thesis

1- Develop an emotion recognition model in real-time using a modified GoogleNet (Inception) algorithm. This system aims to detect emotions based on facial expressions, head position, and eye movements extracted from video data.

2- The research focuses on comparing and emphasizing the distinctions in facial features between individuals diagnosed with Asperger syndrome (AS) and those who are typically developing (TD). By establishing these variations

1.6 Thesis Layout

Chapters of the thesis are organized in the following order:

- **Chapter Two: (Theoretical Background)**

In this chapter different theories of emotions are described to identify the mechanism that controls emotions and understands the relationship between them. In addition, the methodological tools used to measure these emotions are described. Next, Artificial Intelligence, the Convolutional Neural Networks CNN and their equations are presented.

- **Chapter Three: (Proposed System Design)**

This chapter presents a proposed method for Design.

- **Chapter Four: (Results and Discussion)**

This chapter presents the description of the different experiments and discusses the results and evaluations obtained from the implementation of the proposed system.

- **Chapter Five: " Conclusions and Future Work"**, The conclusions and future work will be presented in this chapter.

Chapter Two

Theoretical Background

2.1 Introduction

Different theories of emotions are described in this chapter to identify the mechanism that controls emotions and understands the relationship between them. In addition, the methodological tools used to measure these emotions are described. Next, Machine Learning, the Convolutional neural networks CNN and their equations are presented. Finally, Performance measurement such as accuracy, precision, and recall are also described.

2.2 Emotions

Emotion means the inner and outer feelings of a person. It is difficult to define emotion, because it involves different domains and contexts [27]. Stating that emotions have many definitions, but they agreed on a concept, that emotions are responses to events involving needs, desires, and human interests. Detecting and understanding someone's emotions has grown significantly in importance in recent times. Emotions play a crucial role in interpersonal communication, making it an integral part of our daily lives. Emotion is a multidisciplinary field involving various disciplines such as psychology and computer science. In psychological terms, it refers to a mental state connected to thoughts, feelings, behavioral responses, and a sense of satisfaction or dissatisfaction[28] . People communicate primarily through their positive, negative, or neutral emotional responses. Positive emotions are commonly conveyed through various descriptive words like cheerful, happy, joyful, and excited, while negative emotions include hate, anger, fear, depression, sadness, and others, and so on [29].

Figure 2.1:plutchik's model for Basic Emotions[30]

2.3 Choice of Communication Signs

The diagnosis and forecasting of emotional states rely on interpretations taken from many communication channels, including verbal, non-verbal, and physiological channels. Non-verbal channels include facial expressions, eye look, head movement, posture, and gestures[31]. The verbal channel is audio. Physiological channels are used to acquire emotional data by invasive or non-invasive interventional equipment such as skin conductance, brain wave, and heart rate[32]. Regarding the selection of the most significant communication channel that should be employed for detecting and identifying human emotional state, there are conflicting opinions. The relative significance of the communication channel depends on a variety of elements, such as the nature and qualities of the data sent through the various channels, the judges' interpretations, the kind of message delivered, and the type of affective state

expressed. In the end, researchers in psychology and affective emotions believe facial expressions crucial for displaying human emotion and evaluating human behavior. The human face is a natural source of data for identifying emotions, intentions, and empathy[33]. In addition, with the development of computer vision algorithms, face analysis may be performed using just a camera to collect a live picture stream and no specialized equipment[34]. Consequently, facial expressions together with head motions and eye gazing are a crucial option for extracting emotional states in this thesis.

2.3.1 Facial Expressions

People use non-verbal cues to communicate and express themselves with others. Facial expressions, head movements, and eye gaze are nonverbal signs. In addition to postures, which indicate the body's actions while engaging with others, people employ various channels or signals to communicate non-verbally[35]. Researchers in affective computing and psychologists have focused increasingly on the human face as the visible human element that communicates emotion[36]. Although the human face plays a key part in human communication and emotion recognition[29], head motions may also play a significant role[37]. Expressions on the face are the result of facial muscle contractions; these contractions govern the face's basic features (such as; lips, eyes, eyebrows, nose). Paul Ekman Developed a Facial Action Coding System (FACS) to quantify the facial muscle movements caused by various emotions. Figures (2.2 and 2.3) depict 46 action units (AU) for the upper and lower facial regions, respectively. Some AUS may manifest alone or in conjunction with other emotions. In addition, Ekman has classified a group of six fundamental emotions (1982). Ekman's work still forms the basis of other researchers works in automated and analysis facial expressions and action motion measurements in the human face[12].

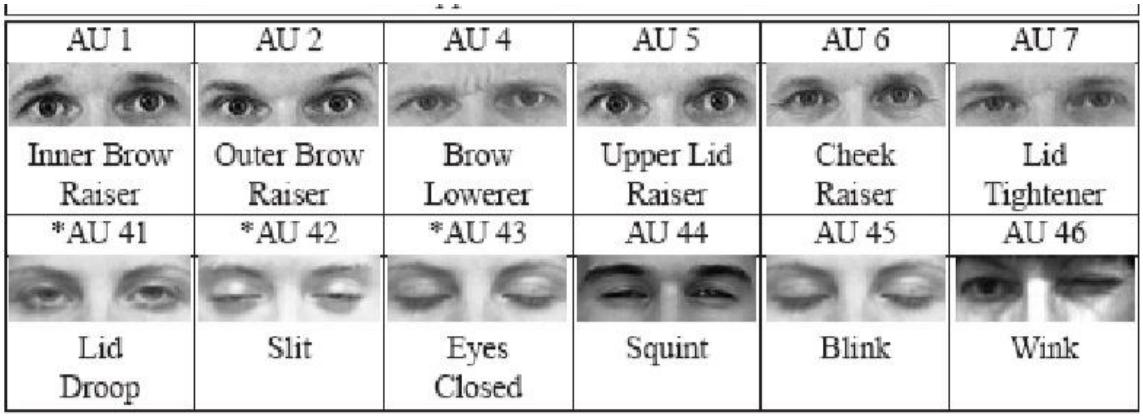


Figure 2.2:Action Units(AUs) For Upper Face[38]

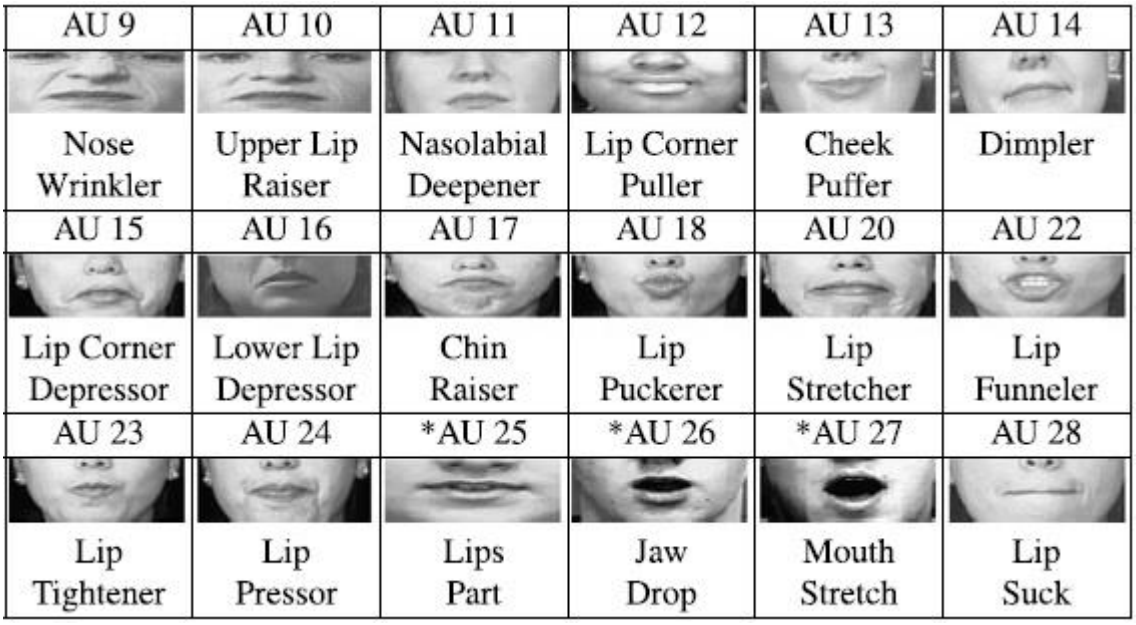


Figure2.3:Action Units(AUs) For Lower Face[38]

2.3.2 Head Pose, Eye Gaze and Emotional State

Head postures are a significant means of non-verbal human communication and hence a critical aspect of comprehending how humans interact with their surroundings. Faces are key to comprehending human behavior. Face perception aids in the comprehension and interpretation of information

processing inside the human mind[39]. Facial features provide an abundance of social information. For a computer to comprehend human behavior, it must infer the human face visual signal with extreme precision, speed, and robustness regardless of occlusion, difficult angles, illumination, etc. The two most significant features of human faces are the expression and head position. Head position orientation may be used to determine the visual focus area of a person. Human emotions assist adaptive responses to environmental constraints. Continuous emotion recognition is crucial for affective computing and human-computer interaction. The majority of available video-based continuous emotion identification algorithms rely on facial expressions. However, in addition to facial expression, additional cues such as head posture and eye gaze are also strongly associated with human emotion, although they have not been thoroughly investigated in continuous emotion identification tasks. On the one hand, head position and eye contact may influence the believability of facial expression traits. In contrast, head posture and eye contact provide emotional cues that are complimentary to facial expression[40].

- **OpenFace Tracker**

  OpenFace, written in C++, is a state-of-the-art tool capable of real-time tracking tasks such as detecting face landmarks, tracking head position and eye fixation. Despite the growing researches in facial behavior analysis, it is challenging to find open-source tools that perform all these tracking tasks, particularly in real time. OpenFace was employed to track the movements of key facial points from live or recorded video streams. The tracker utilizes Conditional Local Neural Fields (CLNF) based on Constrained Local Model (CLM), an algorithm suitable for face and head pose tracking in 3D models with varying appearances and shapes. Furthermore, this algorithm can handle low illumination, head movements, and occlusion[41].

2.4 The Emotional Dataset

Building an emotional dataset is the first step in designing any emotional computer model for emotion prediction. This is a tough stage in the identification of emotional and facial expressions. Different datasets were created for this purpose, and nearly all of them evaluate the six main emotions (Fear, Disgust, Joy, Anticipation and Sadness) with a frontal face view[42]. In addition, the majority of these datasets were produced using emotions portrayed by professional actors. Recently, academics have moved their focus to predicting spontaneous emotions for practical purposes. The study confirmed that there are differences in the look and intensity of spontaneous and performed facial emotions [43].That means the system that was created based on deliberate/acted data cannot be used for intelligent human-machine or human-human interaction. Therefore, there is a need for natural datasets as opposed to artificial datasets that can be utilized in real-world applications, as well as a design methodology based on these datasets. in addition to the difficulty in labeling the intended emotion due to the face's ability to include several emotions in a single instance, constructing natural datasets is a significant problem that necessitates increased work and time. On the basis of the type of data, emotional datasets are often categorized into three types.

- **Acted Data** :in this type, professional actors are coached to convey particular emotions. The majority of automated emotion systems and facial expression identification utilized acted data. These datasets can be easily accessed, they do not challenge ethical considerations, data collection and annotation are simple ,and they may collect a vast quantity of data for each emotion type .However, acted data has less variation of spontaneous data and hardly reflects the real interpretation of individuals behaviors[44].

- **Induced Data:** This type of data is gathered by inducing or eliciting particular behaviors in a controlled environment. In this strategy, participants were exposed to scenarios designed to trigger particular emotions. This sort of dataset has more realistic data than acted data. However, the strength of the emotions created by these approaches is modest, they require control settings, and the participants are aware of the desired emotions[20][45][43].
- **Spontaneous Data** : This type involves the collection of data in a natural and uncontrolled setting. The participants in this category are unaware of the types of feelings they will generate. Several spontaneous datasets, including interaction with machine , human-to-human contact, a dataset for call center voices, and a dataset for videos from websites, were developed [46][47]. These datasets produce natural emotions, but there are ethical issues like information confidentiality, the difficulty in collecting and annotating this data.

  Generally, each dataset has advantages and disadvantages, however choosing a suitable dataset depends on the context that is used to elicit emotions as well as where this data will be used.

  For this thesis, Spontaneous Data was used to build the proposed system because it captures real emotions and reactions , meaning that it reflects the complexity and diversity of emotions in real-world contexts. the dataset contains the basic emotional states to recognize the emotions of individuals with Asperger's syndrome and to identify the differences between the facial expressions of AS and TD when they exhibit the same emotion.

1. Natural-Spontaneous Affective Cognitive Dataset (NACD)

It is Spontaneous data, videos were recorded in a classroom setting under uncontrolled conditions. These conditions include different occlusions (partial obstructions), illumination changes, and background variations. The

purpose of capturing videos with these variations is to create a dataset that reflects real-world scenarios and challenges commonly encountered in educational settings. During the recording process, the students interacted with a computer and engaged in the learning process using educational games. The dataset captures their natural and spontaneous emotional responses during this interactive learning experience. The NACD dataset provides a valuable resource for studying and understanding the affective cognitive processes of both typical development and autistic students[48].

2.5 Machine learning

Machine learning (ML) is a subset of artificial intelligence (AI) that focuses on the development of algorithms and statistical models that enable computer systems to improve their performance on a specific task through experience or data. The central idea behind machine learning is to allow computers to learn from data and make predictions or decisions without being explicitly programmed for each specific task. Machine learning systems work by analyzing and processing large amounts of data to identify patterns, relationships, and insights [49].

2.5.1 Artificial Neural Network (ANN)

Artificial neural networks are computational approaches that mimic how the human brain functions for a certain job via the vast manipulation of fundamental units that are tightly coupled in groups. Single-layer neural networks, alternatively known as single-layer neural networks, consist of only one input layer and one output layer. On the contrary, multi-layer neural networks are neural networks that include a hidden layer. If a multi-layer neural network has only one hidden layer, it is referred to as a shallow neural network.

Conversely, a deep neural network is characterized by having more than two hidden layers. This distinction is illustrated in Figure (2.5)[50] .

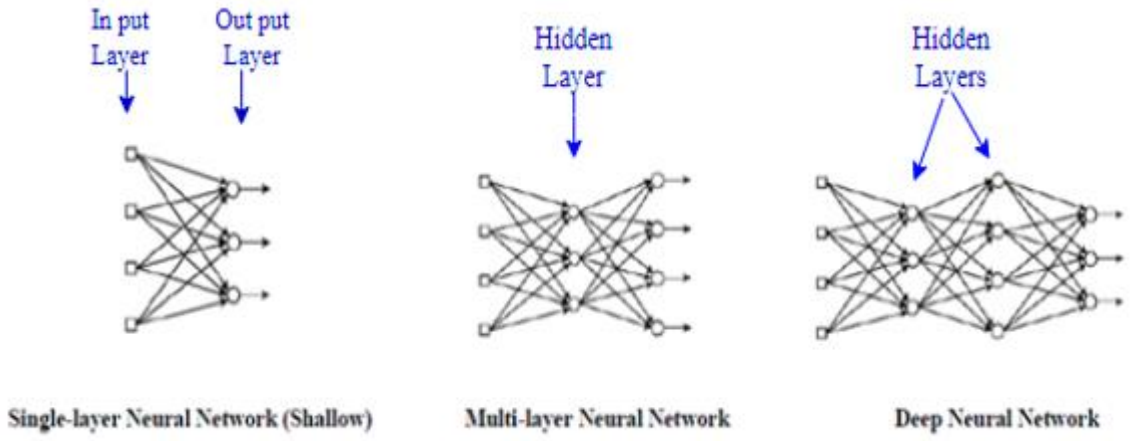


Figure 2.4:The Types of the NN Based on Architecture of Layer[50].

MLP (Multi-Layer Perceptron) is one of the most often used ANN models.

Recurrent neural network(RNN), which is recurrent and CNN which is convolutional neural networks, are two other popular ANN models[51].

A- The Activation Functions

An activation function is a type of function employed to carry out mathematical operations on the resulting output. The specifics of the issue that the network is attempting to address should guide the selection of the appropriate activation function. Sigmoid, Hyperbolic Tangent (tanh) and Rectified linear units Relu, are depicted in Equations (2.1), (2.2), and (2.3), respectively. These three activation functions are the ones that are utilized as activation functions in neural networks the majority of the time.

$$F(x) = \frac{1}{1 + e^{-x}} \quad (2.1)$$

$$F(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}} \quad (2.2)$$

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases} \quad (2.3)$$

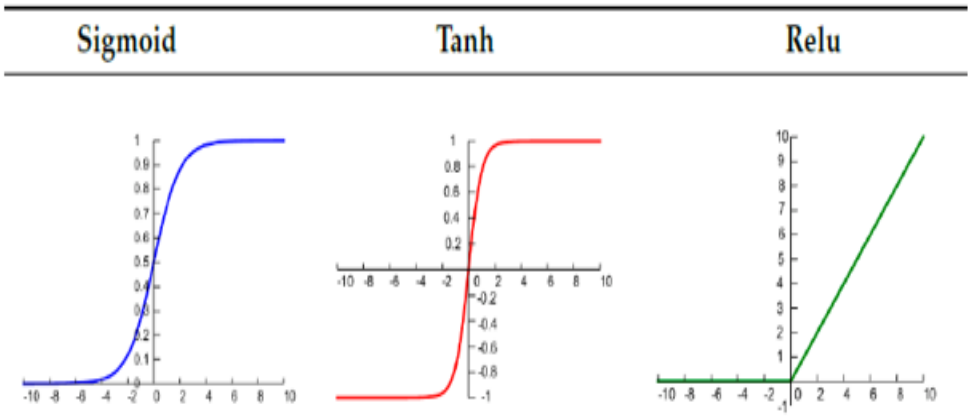


Figure 2.5:Illustrates Activation Functions[52].

Figure 2.5 demonstrates that the sigmoid and Tanh functions are comparable in terms of the range of the input value x, which lies between -1 and +1, and the range of the output value f (x), which lies between 0 and 1 for the sigmoid function and between -1 and 1 for the tanh functions, respectively.

The vanishing gradient issue occurs when the input value of x is huge but the output values of f(x) for the two functions are virtually zero. This is because the vanishing gradient problem is an inherent disadvantage for both functions. In order to find a solution to this issue, the ReLU activation function will only take into account the positive signal and will disregard the negative signal. It has an excellent fitting ability and sparsity, which considerably enhances the calculation performance and may successfully avoid a vanishing gradient. This is because it has a high fitting ability and sparsity.

It is also possible to use an output layer activation function called a softmax to evaluate probability and do multi-class classifications. The formula (2.4) below is used to apply the softmax function:

$$\boldsymbol{Soft\ Max(Xi)} \quad = \frac{\mathbf{e^{zj}}}{\sum_{k=1}^{K} \mathbf{e^{zk}}} \quad \boldsymbol{for\ j = 1. \ldots . K} \qquad \textbf{(2.4)}$$

Where *K* is the number of classes, $\boldsymbol{zj}$ is the production corresponding to class j[52].

2.5.2 Deep Learning

It is possible for computers to learn to represent data at numerous levels of abstraction using deep learning. Speech recognition, optical object identification, object detection, and a host of other fields have benefited immensely from these techniques. Deep hidden layers of early neural networks were difficult to train, which reduced their performance. This challenge was addressed using Deep Learning. Convolutional and recurrent networks have made substantial improvements in processing pictures, video, voice, respectively, whereas recurrent networks have focused on sequential data like text and speech[53].

2.5.2.1 Deep Convolution Neural Network (DCNN)

A deep neural network type specifically created for image recognition applications is the convolutional neural network (CNN). It is also referred to as ConvNets or deep CNN (DCNN). It is important to note that CNN differs greatly in several ways from traditional networks, including the fact that CNN uses images as its primary form of input rather than features, and hidden layer neurons are connected to a portion of the neighbouring layer neurons while in the neural network each neuron is connected to all the neurons in the neighboring hidden layer. This differences allows CNN to learn from images

more effectively than traditional networks. There are many layers in DCNN, and it is meant to mimic how the brain recognizes pictures, making it a more than simply a deep neural network. Instead of manually constructing features, ConvNet can extract features by using several locally linked layers automatically[53][54]. Experts designe the feature extractor prior to the development of ConvNet. As a consequence, it required a great deal of time and money, and the results were inconsistent. As seen in Figure 2.6, there is a significant difference between ML and ConvNet.

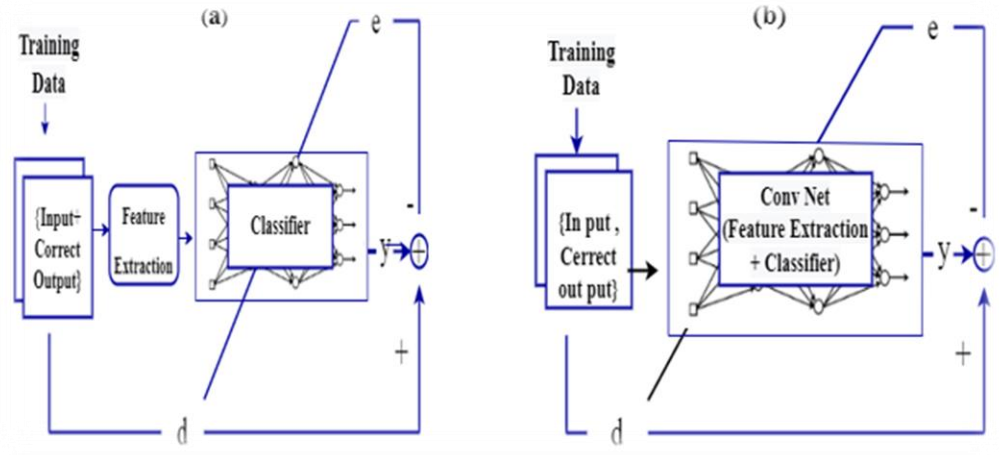


Figure 2.6:(a) Machine Learning with Using Hand-Designed Features; (b) ConvNet with Multi-Layers[51].

A convolutional neural network (CNN) comprises two primary components. The initial component is composed of convolutional and pooling layers, which are employed to extract features from input images and analyze these features. The subsequent component is a sequence of fully connected layers, responsible for predicting the most appropriate classification for the input image. This structure is illustrated in Figure (2.7).

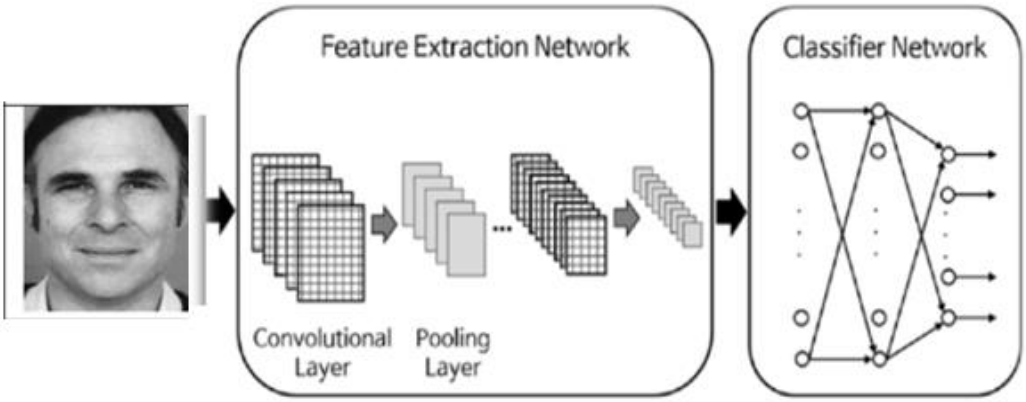


Figure 2.7: Typical Architecture of ConvNet [51].

The fundamental structure of a CNN encompasses three architectural principles that endow it with robustness across various domains, such as image processing, pattern recognition, speech recognition, and Natural Language Processing (NLP).

The first principle is applied in both convolution and pooling layers, wherein each neuron receives input from a small region known as the local receptive field. This receptive field size is equivalent to the dimensions of the convolution filter. By adopting this local connectivity scheme, the trained CNN exhibits heightened responsiveness in capturing local dependencies and extracting distinct features within the input image, such as edges, ridges, curves, and so on[54].

The second principle is to reduce their complexity by using the strategy of sharing parameters (weights) which are applied by the convolution layer. a consistent convolution filter is employed to identify a specific feature. Subsequently, non-linear downsampling is applied in the pooling layers to reduce the spatial dimensions of the sample obtained from the convolution layer. Additionally, the number of model parameters is restricted. These

characteristics collectively enhance the stability and efficiency of Convolutional Neural Networks (CNNs) during their operations[48]. four distinct layers make up the CNN's standard architecture and any Convolutional Neural Network model is built from these different layers:

1. Convolutional layer.
2. The pooling layer (or Sub Sampling layer).
3. The flaten layer
4. Fully Connected Layer (Classification layer).

A. Convolution Layer

The convolutional layer is essential in constructing a convolutional neural network model. It performs a large number of computations. Its primary role is to execute a convolution operation on the input data to extract features. (The convolutional layer consists of three matrices (input, filters, and results). The input matrix includes the input data, converted into a two- or three-dimensional matrix. The filter matrix is also called the kernel, where it is small in spatial dimensions but can spread to all directions. Finally, the resulting matrix is the result of moving the filter over the input image to calculate the point product called the feature map. The activation maps are then concatenated to create a single convolutional layer's final output, which serves as input data for the next layer[50].

Multiple convolution filters are applied to a single entry, and each value of the filter array is provided by default. However, it should be noted that filter values are different from one filter to another to provide unique properties for each feature map as shown in Figure (2.8) [51].

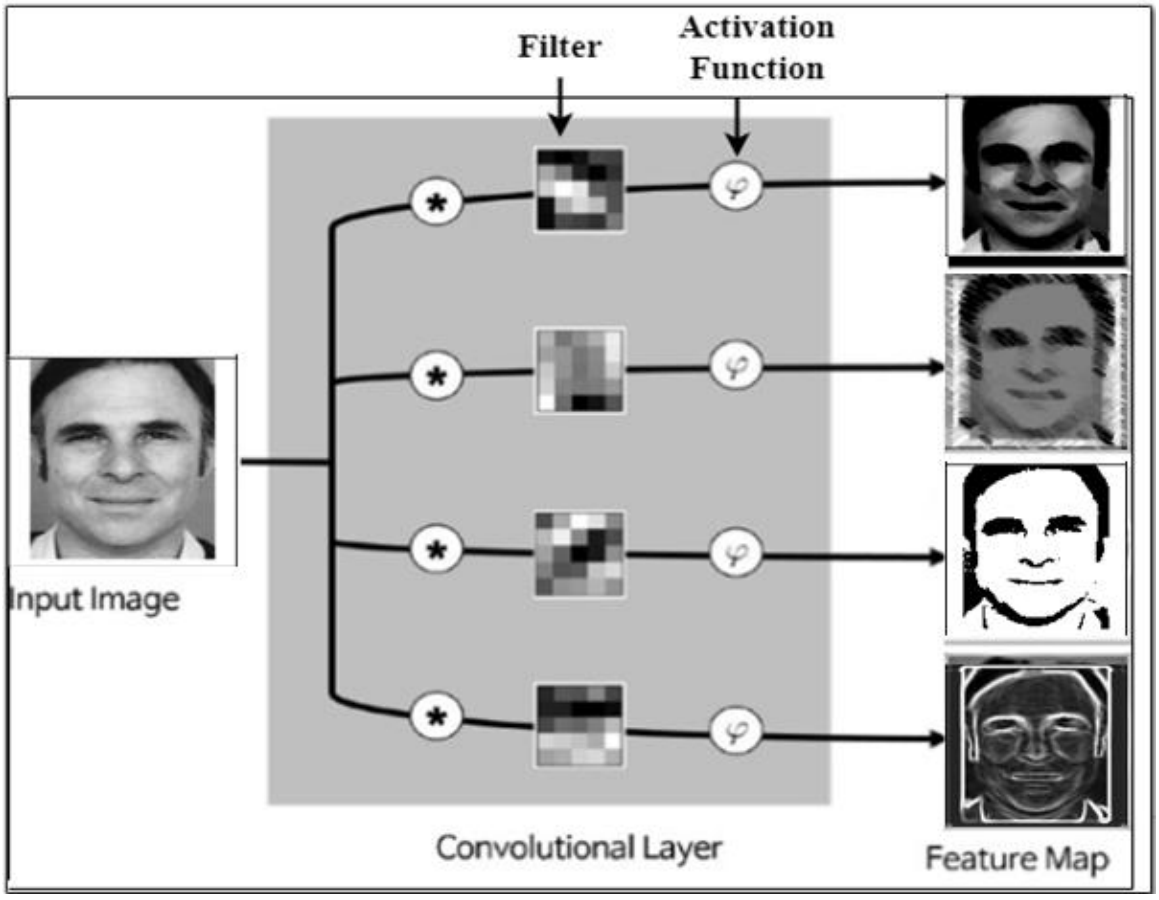


Figure 2.8: Convolution Operation [51].

The following Equation (2.5) describes a general convolutional layers process for images $(i.j)$

$$z_{ij} = (x * k)_{(i.j)} = \sum_{l=1}^{k1} \sum_{a=1}^{k2} \sum_{b=1}^{c} x_{(i+l.j+a.b)} k_{(l.a.b)} \qquad \textbf{(2.5)}$$

The feature map Z is produced by the convolution process between the input X and the kernel weights K in the convolutional layer l in the preceding equation. Here, $K1$ represents the vertical dimension of the kernel, $k2$ represents the horizontal dimension of the kernel, and C denotes the number of channels.

There are three hyper-parameters to set when designing the convolution layer: depth, stride, and zero- padding[51].

1. **Depth**: The depth signifies the quantity of filters employed during the convolution process. The activation map' depth' would equal three if the original image were convoluted using three filters.
2. **Stride:** refers to how many steps in a convolution operation we transfer in each step. If the stride is 1, the kernel will travel by one step. If it equals 2, the kernel will advance by two steps. The size of the function maps decreases as the number of steps rises.
3. **Zero-padding**: Increases the number of rows and columns by 2 by surrounding the input matrix with "0" values; if p = 1, It is possible to make larger padding. If p = 2, then the number of rows and columns increases by 4, and so on. When these parameters are used, this changes the spatial dimensional output of the convolution layers. Formula (2.6) gives details of this change [41].

$$\boldsymbol{output = floor\left(\frac{input\ image\ size - kernel\ size + 2 * Padding}{Stride} + 1\right)} \qquad \textbf{(2.6)}$$

There are two types of convolution: valid convolution and as a result of this type of convolution, no padding is used, and the array size will gradually decrease. The second kind is zero or "same," which employs a padding mechanism, and as a result, the array's size remains constant before and after the convolution[55]. Figures (2.9) and (2.10) show the two types of convolution processes. It is important to mention that in Figure (2.10), the output feature map retains the same input dimensions of 5 * 5. However, in Figure (2.9), the output feature map is reduced to 3 * 3.

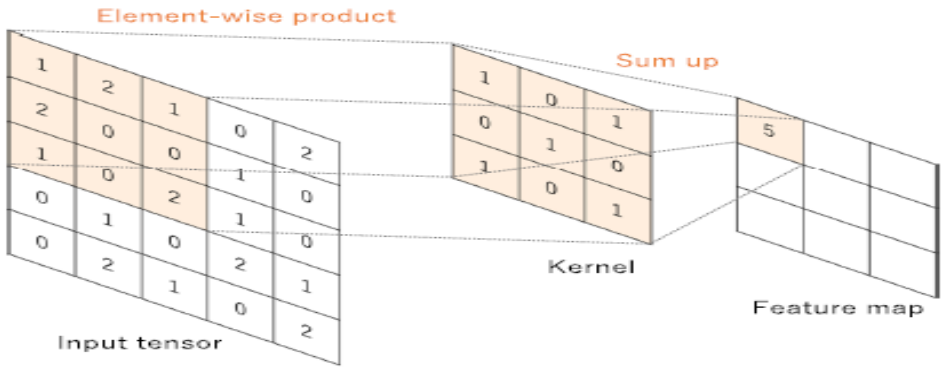


Figure 2.9: Convolution Operation Without Padding, Using a 3x3 Kernel Size[56].

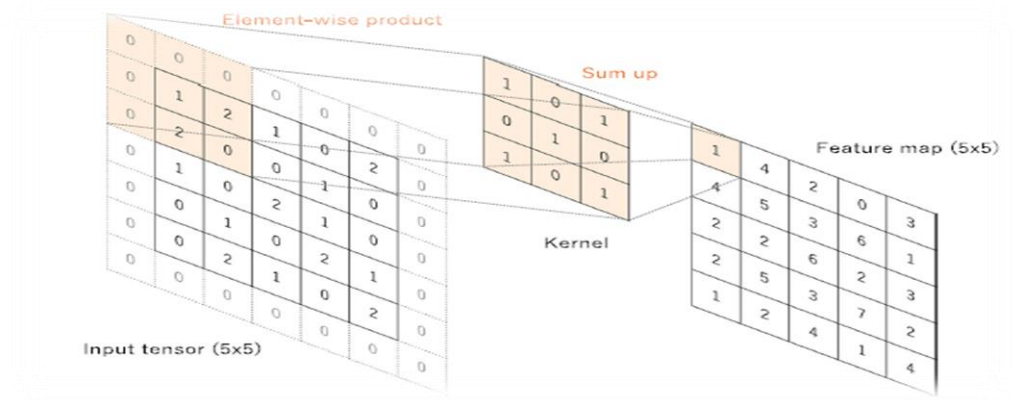


Figure 2.10: convolution operation with a 3 × 3 kernel and 1Padding value [56].

B. Pooling Layer

It is The technique of down sampling neighbouring pixels into a single pixel, which captures more robust information, the output dimensions of the previous layer are lowered. In addition, the two types of pooling are maximum pooling and average pooling. In first type finds the maximum pixel value, whereas the second retrieves the average pixel value from a feature. Figure (2.11) depicts the workings of both pooling systems. The max-pooling down sampling process is described as in Equation (2.7).

$$si = max\ hi \qquad i \in Rj \qquad \textbf{(2.7)}$$

Where h represents some pixel in the window (or sub-region) Rj from the rectified Activation Maps

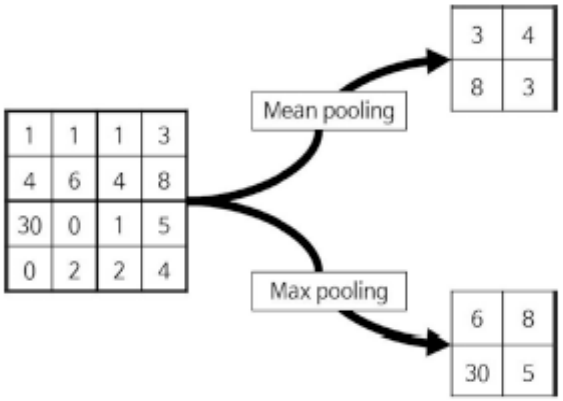


Figure 2.11: The Pooling Layer [51].

C. Flatten

After obtaining the feature map (2D representation of the image) from the multiple convolutional stacks, it is converted to a vector as it passes through the flatten layer and then pass it to the fully connected layer [54], as shown in Equation (2.8)

$$y_{ijk} = \frac{e^{xij}}{\sum_{T=1}^{D} e^{ijt}} \quad \textbf{(2.8)}$$

D. Fully Connected Layer

Once the previous layers have been repeated several times, the data is combined into a flat vector and entered into the neural convolution network's fully connected layer. There are complete connections between all nodes in the preceding layer, which is represented by the neurons in the Fully connected layer. The purpose of this layer is to consolidate the weights of the features obtained from the preceding layers and provide the probability of each class[56]. In Equation (2.9) describe Fully connected layer.

$$Fc_p^{(l+1)} = \sum jC_j^l \, W_{jp}^l + \alpha_p^l \qquad (2.9)$$

W_{jp}^l is a matrix weight between node j of layer L and node p of layer (L+1), where *cjl*is is the data of node j in layer L.

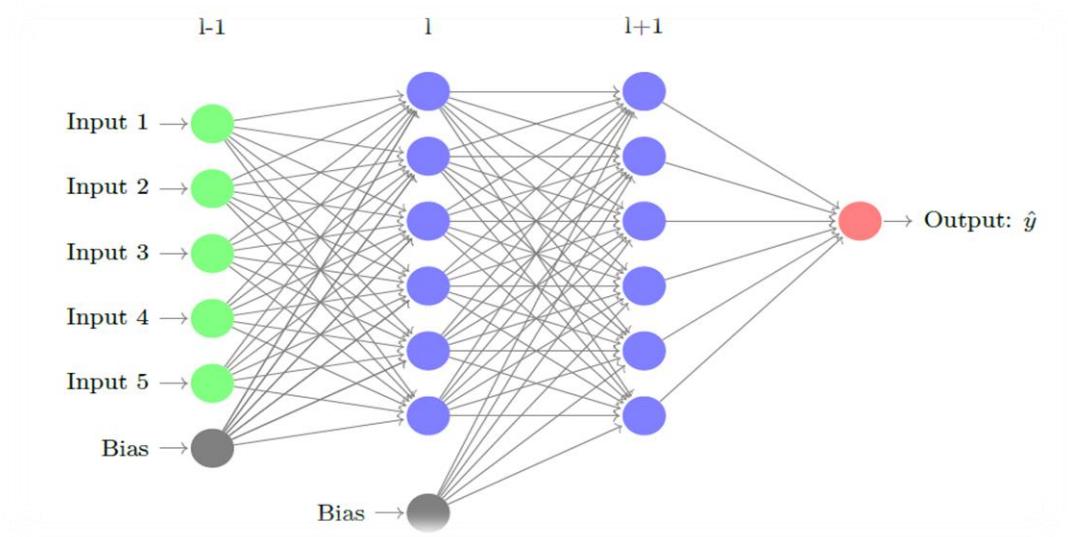


Figure 2.12: fully connected layer[55].

- **GoogleNet**

GoogleNet, also referred to as Inception v1, is a sophisticated convolutional neural network design created by researchers at Google to accomplish image classification objectives. This network achieved remarkable success by securing the top position in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2014, showcasing an impressive top-5 error rate of 6.67%[57].

The network is made up of a series of convolutional layers, pooling layers, and fully connected layers, as shown in figure (2.13). To begin, the input images undergo standard data augmentation methods like random cropping and horizontal flipping. Afterwards, the images are passed through several convolutional layers, with each layer being followed by a rectified linear unit (ReLU) activation function and a max-pooling layer[58]. The output of the last

pooling layer is fed into several inception modules, which are connected to a global average pooling layer and a softmax layer for classification.

To train the network, the researchers used stochastic gradient descent (SGD) with momentum as the optimization algorithm and cross-entropy loss as the objective function. They also used a technique called dropout to prevent overfitting. The network was trained using the ImageNet dataset, which includes 1.2 million labeled images in 1,000 categories.

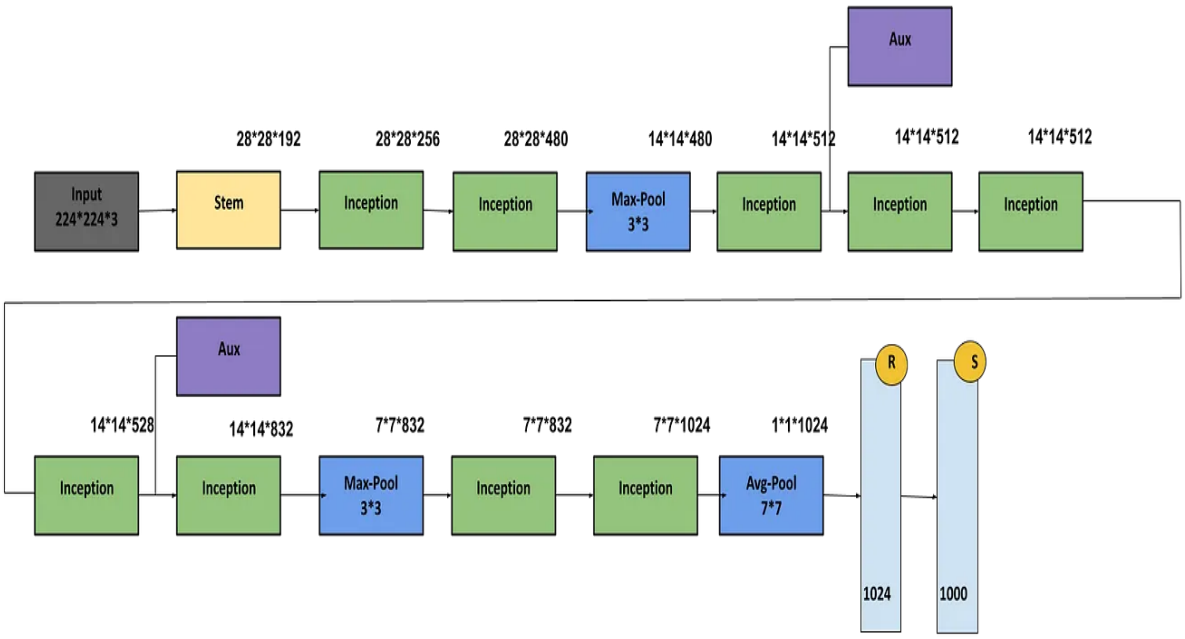


Figure 2.13:GoogleNet architecture[57]

2.5.3 Optimization Algorithm

Most of the training methods applied are based on the backpropagation algorithm by propagating the errors backwards from the output nodes and updating the weights of each layer with the gradient descent optimization algorithms. In order to further improve training results, several novel algorithms have been presented. Adam is an optimization technique that makes use of iterative weight updates based on training data in place of the conventional

stochastic gradient descent method. The term "Adam" is coined from "adaptive moment estimation." This approach combines the concepts of Root Mean Square Propagation (RMSProp) and momentum techniques. By calculating the first and second moments of gradients, it determines adaptive learning rates for different parameters. Adam's optimizer is widely recognized and widely utilized in the realm of adaptive learning algorithms [59].

$$w_{t+1} = w_t - \frac{\alpha}{\sqrt{v_t + e}} \cdot \hat{m_t} \quad \textbf{(2.10)}$$

$$\hat{\boldsymbol{m_t}} = \frac{\boldsymbol{m_t}}{\boldsymbol{1 - \beta_1^t}} \quad \textbf{(2.11)}$$

$$\hat{\boldsymbol{v_t}} = \frac{\boldsymbol{v_t}}{\boldsymbol{1 - \beta_2^t}} \quad \textbf{(2.12)}$$

$$\boldsymbol{m_t = \beta_1 m_{t-1} + (1 - \beta_1) \frac{\partial L}{\partial w_t}} \quad \textbf{(2.13)}$$

$$\boldsymbol{v_t = \beta_2 v_{t-1} + (1 - \beta_2)[\frac{\partial L}{\partial w_t}]} \quad \textbf{(2.14)}$$

M and V represent moving averages, and ß1 and ß1 are decay rates used for estimating moments. The standard values for ß1 and ß1 exponential decay rates are 0.9 and 0.999, respectively[55].

2.5.4 Epoch in CNN Training

The number of epochs serves as a hyper parameter determining the frequency at which the learning algorithm processes the complete training dataset. Each epoch signifies an iteration where every sample in the dataset has an opportunity to modify the internal model parameters[55].

2.5.5 Batch in Training

The Batch size defines the number of samples that operate in the internal model before changing the parameters. A Batch is a process of iterating one or more samples to make a prediction. The predictions are compared with expected

output variables at the end of the Batch, and an error is measured. Then, the update algorithm uses this error to enhance the model. There are three types of Batches, as explained in the following:

- Batch gradient descent refers to the process of using all of the training samples to build a single Batch.
- Stochastic gradient descent (SGD) if the Batch is no larger than a single training sample.
- Mini-Batch gradient descent is employed when the size of the Batch is greater than a single sample but smaller than the size of the training dataset.

  The mini-Batch is most widely used in deep learning training technique because it prevents the model's sluggish convergence and lowers gradient disappearance and convergence instability [60][61].

2.5.6 Learning Rate

The pace at which a model learns is a critical parameter in the training process. Learning rate calculates the amount of parameter adjustments and minimizes the disparity between the predicted and actual outputs. If the learning rate is set too high, the training speed will be faster, but the model may struggle to converge to the minimum. Conversely, if the learning rate is very low, the model's training accuracy may improve, but the convergence rate will be slow, increasing the risk of getting stuck in a local optimum. Learning rate values consider a hyperparameter and can be adjusted in a neural network with small and positive values, often between 0.0 and 1.0 [62]

2.6 Deep Learning Training Skills

To enhance results and prevent model overfitting, deep learning training techniques are commonly employed during the model training phase. Among

these techniques, batch normalization and dropout have been identified as highly effective deep learning algorithms[63]:

2.6.1 Batch Normalization

In deep learning, batch normalization is a typical training method. The batch normalization approach maintains sample distribution features inside a single layer and eliminates the distribution gap between layers. Equation (2.15) presents the definition of the batch normalization function.

$$\mathbf{X'}\boldsymbol{i} = \boldsymbol{x} = \frac{\mathbf{Xi} - \mathbf{X^-}}{\mathbf{Xvar}} \qquad \textbf{(2.15)}$$

Here, x represents the mean value of n input data points x, while $\boldsymbol{Xvar}$ denotes the variance value of the same set of n input data points x [64].

2.6.2 Dropout

Dropout is a simple and powerful method employed to mitigate overfitting and enhance the overall performance of a model during training. This technique involves randomly disregarding portions of hidden neurons in designated hidden layers during training, with the probability of neglect determined for each neuron (selected hyperparameter p where $0 \leq p \leq 1$) at each training stage, as shown in Figure (2.14) [55] .

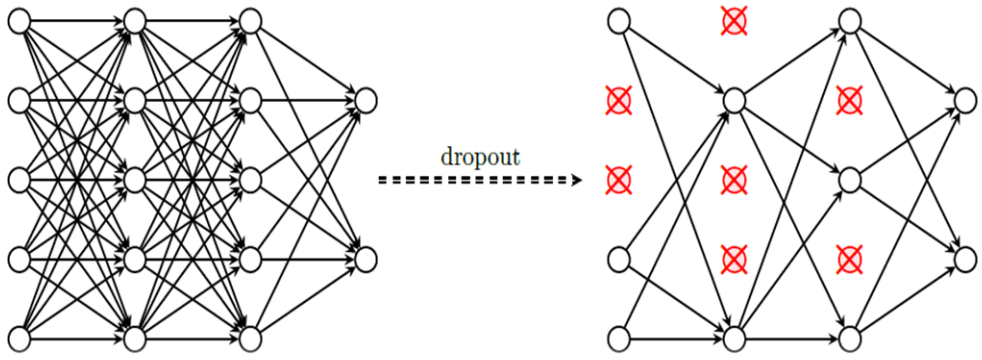


Figure 2.14: shows a Dropout[55].

2.7 Performance Metrics

The diagnostic proposed model evaluates its performance through well-known metrics, including accuracy, precision, and recall, Figure (2.15).

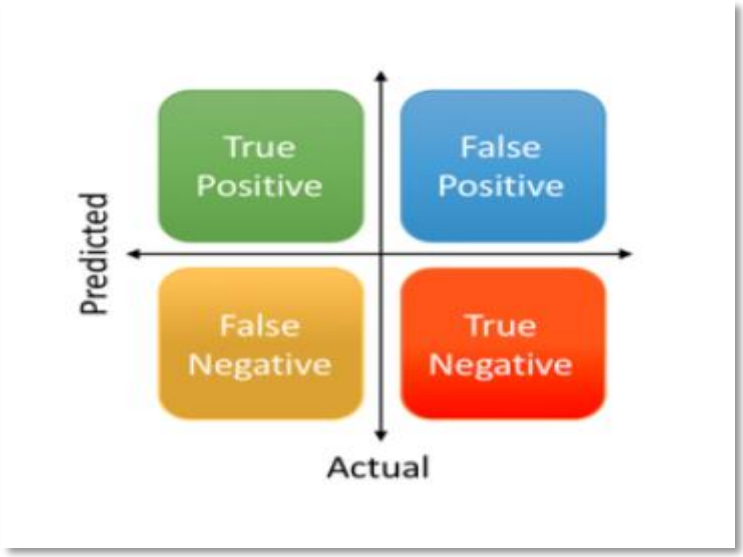


Figure 2.15: displays the Confusion Matrix for the Classifier System.

The performance matrixes need initially calculate a set of parameters after applying the test operation to the collection of images[54]. These parameters are:

- **True Positive** (**TP**): refers to the instances where the model correctly predicts a positive class, aligning with the actual true class.
- **True Negative** (**TN**): denotes cases where the model accurately predicts a negative class, matching the actual false class.
- **False Positive** (**FP**): signifies instances where the model incorrectly predicts a positive class, despite the actual class being false.
- **False Negative (FN)**: The model is predicated improperly, therefore the model is false and the real class is true.

2.7.1 Accuracy

The most trustworthy performance metric is accuracy, which is computed as the proportion of properly predicted observations to all observations. It is not required to assume that a model is operating at its best when it reaches high accuracy. Because the false positive and false negative values are comparable when the dates are symmetrical, other factors must be examined to test the findings. To attain accuracy, apply equation (2.16) [54].

$$\boldsymbol{Accuracy} = \left[\frac{\boldsymbol{TP} + \boldsymbol{TN}}{\boldsymbol{TP} + \boldsymbol{TN} + \boldsymbol{FP} + \boldsymbol{FN}} \times \mathbf{100}\right] \quad \textbf{(2.16)}$$

2.7.2 Precision

The ratio of successfully predicted positive observations to all correctly predicted positive observations may be used to measure precision using Equation (2.17) [65].

$$\boldsymbol{Precision} = \left[\frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \times \mathbf{100}\right] \quad \textbf{(2.17)}$$

2.7.3 Recall

Recall or sensitivity can be obtained by using Equation (2.1), which calculates the ratio of correctly predicted positive observations to the total number of observations in the actual class being true[66].

$$\boldsymbol{Recall} = \left[\frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \times \mathbf{100}\right] \quad \textbf{(2.18)}$$

Chapter Three

THE PROPOSED SYSTEM

3.1 Introduction

This chapter focuses on utilizing deep learning for recognizing emotional states. The chapter will present the aspects and algorithms that are used for building the model. The first section includes data model preparation as it describes the taken approach to prepare the training data. Next, the second section discusses the used layers used to construct the learning model and its structure. Finally, evaluation of model performance towards unseen data by using different metrics.

3.2 Model Design of the proposed Model

The proposed model comprises of several phases to accomplish its task. The model is a supervised learning model, which means it need to train on a related dataset and validate then after obtaining the final model it should evaluate the unseen data to measure the level of generality that we gain and asses the goodness of the model to work with real data.

The overall structure of the system, as depicted in Figure 3.1, involves several phases in designing the model. Below is an explanation of the steps:

1. Face Tracking and Detection: This initial stage involves detecting and tracking faces in in-plane and out-of-plane input video frames using openface tracker. Face detection algorithms are employed to identify and locate the faces of individuals in the video.

2. Pre-processing: In this stage, the data obtained from the tracked faces is pre-processed to ensure its quality and suitability for analysis. Data cleaning techniques are applied to remove noise or outliers, and the data is standardized to ensure consistency and comparability.

3. Model Building using Deep Learning: In the Next stage, a model for emotion recognition is constructed using deep learning techniques, specifically a deep convolutional neural network (CNN) architecture such as GoogLeNet. These

algorithms are chosen for their capability to effectively extract relevant features and recognize emotions from visual inputs.

4. Evaluation of Model Results: In the fourth stage, the efficacy of the emotion recognition model is evaluated. The testing dataset is utilized to assess the accuracy of the model's predictions.

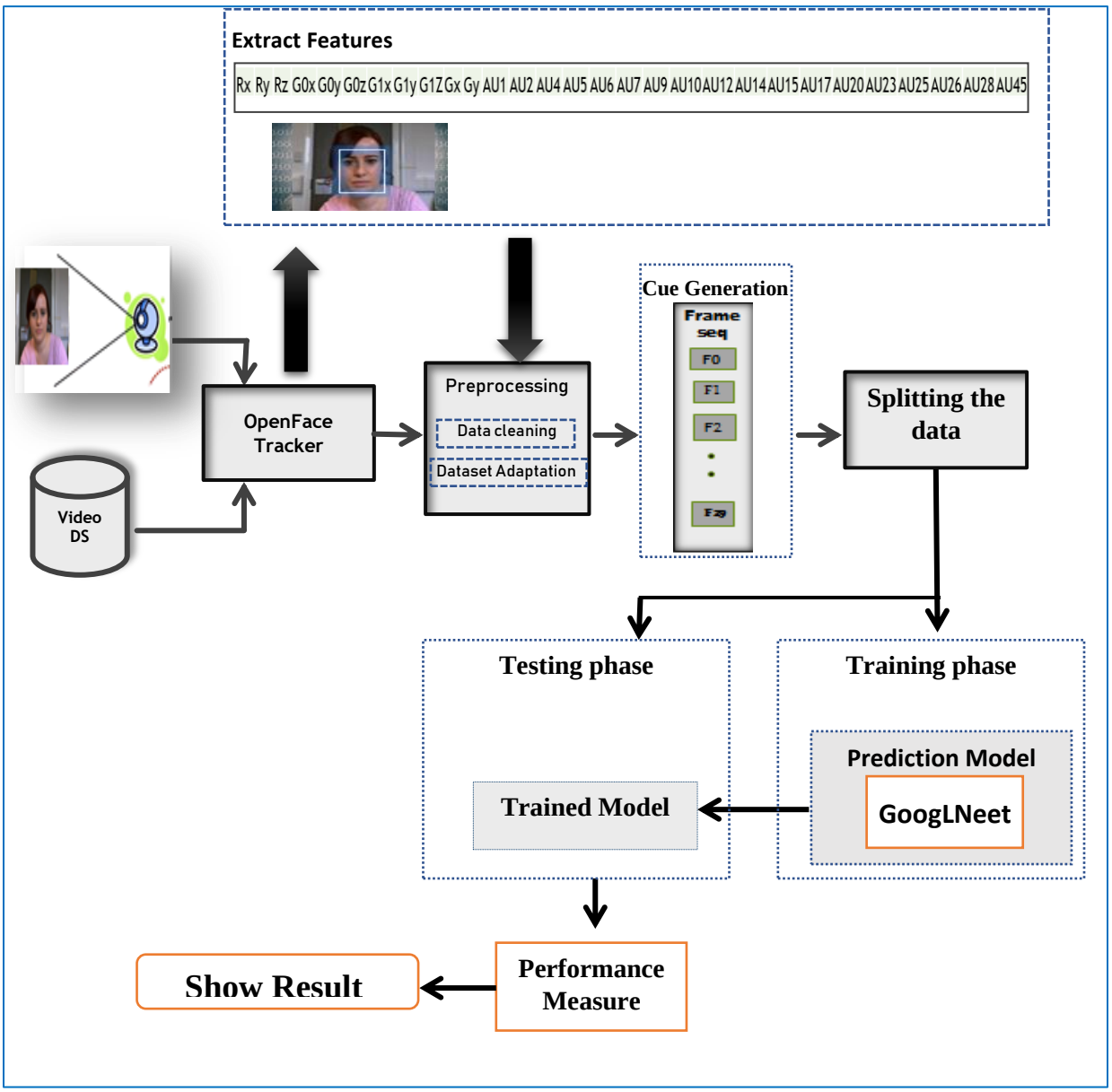


Figure 3.1: Overall Structure of the Proposed Model Lifecycle.

3.2.1 Face Tracking and Detection

The dataset used in this work is called the natural-spontaneous affective cognitive dataset (NACD) [48]. It consists of two folders: one for typical development students (TD) and another for autistic students (ASD). Each folder contains various video clips, and each video clip was labelled with a specific emotion that corresponds to the content of that video.

which were not suitable for the research purposes. To extract relevant information, head ,gaze, and face features were obtained from the dataset video clips using the face recognizer and tracker software, OpenFace[41]. OpenFace, written in C++, is a state-of-the-art tool capable of real-time tracking tasks such as detecting face landmarks, tracking head position and eye fixation. Although the software generates a substantial number of tracking feature points, not all of them were relevant to the research. Therefore, the tracker was modified base on this work need.

The output features from each video clip in the dataset were saved in a comma-separated values (CSV) file. The processing of video tracking is summarized in Figure (3.2). Each structure of the video tracking features contains values about an individual's eye position and gaze, head position relative to the screen and orientation, as well as facial movement units extracted from the learned models.

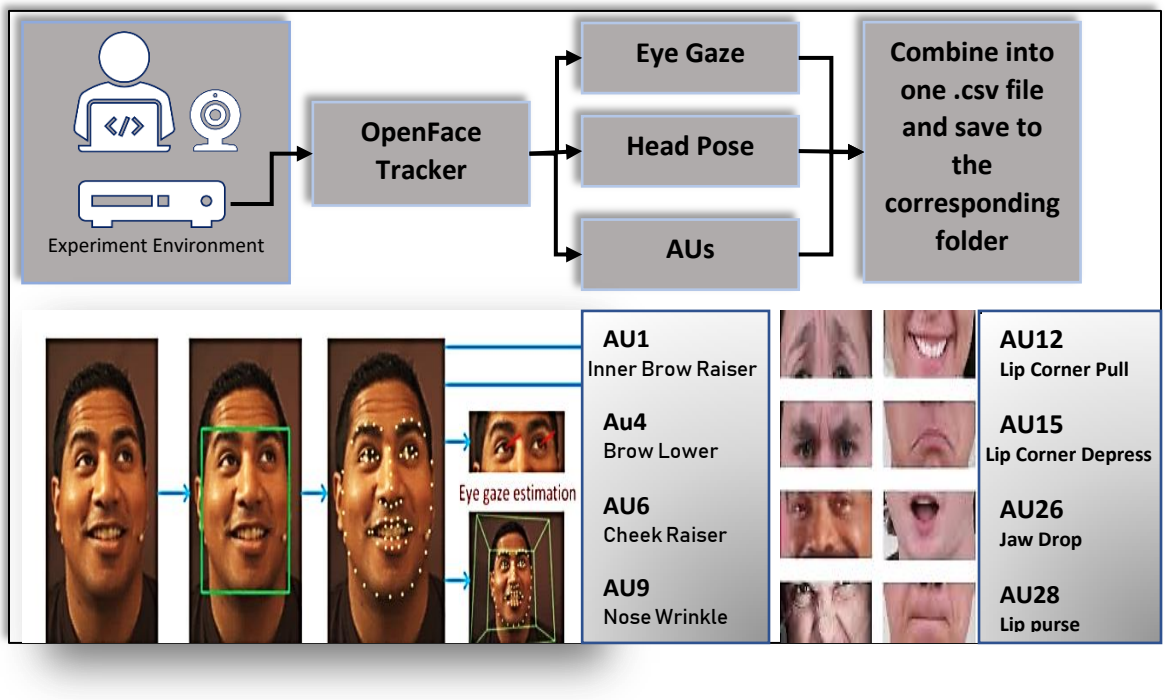


Figure 3.2:Collecting Features Based on OpenFace Tracker Process

3.2.2 Data Pre-processing

Before data using for model prediction, the data should be processed to remove undesired values. Therefore, the next step is data preprocessing. In this step further processes was done like data cleaning, outlier detection, and missing frames values. Then, the data which results from this step will be used as input for emotional model.

A. Data Cleaning

The data collected from the previous step need to be cleaned. Generally, data cleaning is a crucial step in data analysis, and it involves identifying and correcting errors, inconsistencies, and missing values in a dataset.

The type of tasks applicable for this research are correcting errors and standardizing the collected data .

The OpenFace produces results data filtered from outliers and missing data. the only thing that needs to be cleaned in the resulted tracking data is the validity of the processed frame. The validation indicator in the features data file is the "success" filed which is a binary filed ("1" means frame is captured and processed correctly, and "0" otherwise). Frames with success value equal to "0" is eliminated from the training dataset.

B. Standardizing Data

The standardization step in dataset analysis in this research include the transformation of the captured features data to a file that contains only the features of interest for the research. The total number of captured features that recorded in the raw feature file are 334 features. The resulted cleaned features file includes 33 features (frame, timestamp, and success as the frame description data, 6 eye gaze features, 6 head pose features and 18 AUs values). This step generates smaller

sized files that can minimize the time to retrieve and process files during cues generation step.

In this thesis, all phases of model development, focusing on two types of data: individuals with Asperger's syndrome (AS) and typically developing individuals (TD). The primary aim and contribution of this thesis are to identify and understand the differences between AS and TD. To achieve this objective, two separate models were constructed and trained: one for AS and the other for TD. Each model was built using the specific dataset collected from the respective group. The training dataset for each model consisted of processed features files obtained from the initial dataset.

Developing separate models for AS and TD can aid the differences of characteristics in emotional expressions between these two groups. This approach enables a comparative analysis that can provide insights into the unique emotional profiles expressed by individuals with AS compared to typically developing individuals.

3.3 Emotion Cue Generation

To generate an emotion cue, 30 frames (rows) were taken from the tracking file and the features stacked to three categories, gaze, pose, and AUs. Each of these categories arrange in a moving window elapsed rows, the difference between each successive row is 1/30 of a second. This value is chosen based on empirical evidence to simplify presentation and visualization. Additionally, it provides a reasonable latency time for predicting the first label when displaying cues in an online process. The result from this step is three files representing the cue of the same point of time, algorithm (3.1) describe this step and outlined in figure (3.3). To build a single cue consists of a multivariate-temporal time-series with a fixed

length of 30 observations. Each observation contains 30 features. The cue vector has a cue length (c) of 30 and a total number of features (n) equal to the sum of the features in pose, gaze, and AUs, which amounts to 30 (6 features from pose, 6 features from gaze, and 18 features from AUs) (algorithm 3.2) .

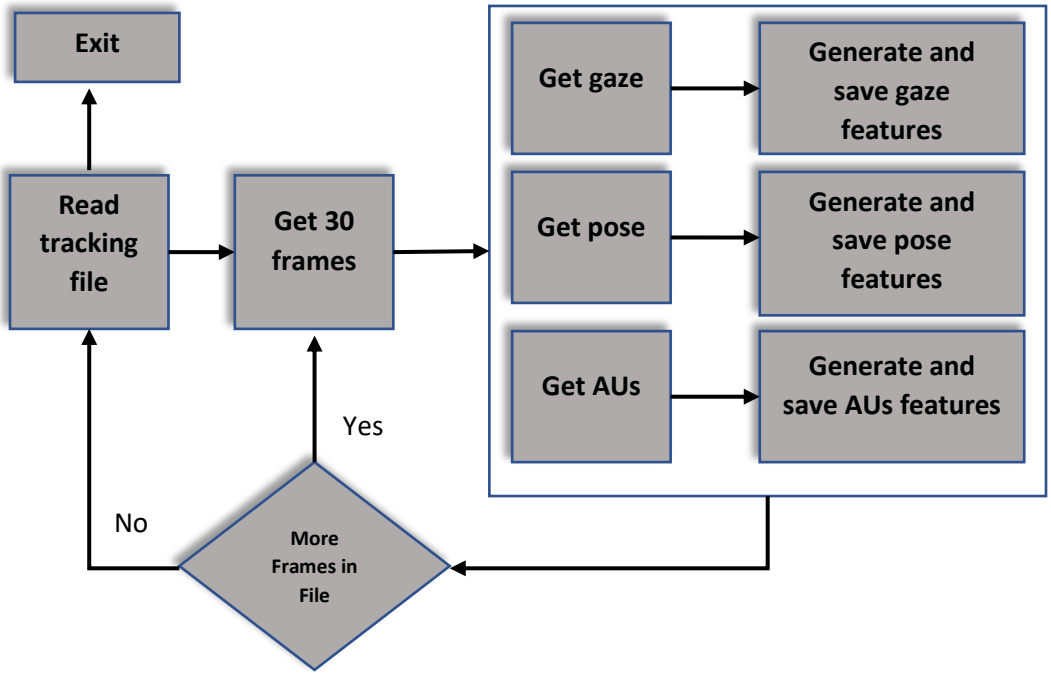


Figure 3.3: Emotion Cues Generation

Algorithm3. 1: MultiDataGenerator:

INPUT: ds_root, list_IDs, labels, dim, batch_size=32, n_channels=1, n_classes=10, shuffle=True

OUTPUT: data generator class

BEGIN

1. Initialize the MultiDataGenerator with the following parameters:
   1.1. ds_root: the root directory of the data
   1.2. list_IDs: a list of IDs of the data samples
   1.3. labels: a dictionary of labels for each sample
   1.4. dim: the dimension of the input data
   1.5. batch_size: the number of samples per batch
   1.6. n_channels: the number of input data channels
   1.7. n_classes: the number of classes in the data
2. get the number of batches per epoch
3. Generates one batch of data:
   3.1. Generate indexes of the batch
   3.2. Generate the corresponding input data and label
   3.3. Update the batches indexes
4. IF epochs number not reached GOTO step 3
5. Generate the results data:
   5.1. Initialize the input data and labels with the given dimensions
   5.2. Generate data and labels for each sample file
6. Return the input data and labels

END

Algorithm 3.2: Generate_cues

INPUT: ds_filenames: names of dataset files, cues_labels: dataset emotions labels

OUTPUT: the resulted dataset cues

BEGIN

1. Read dataset files and labels
2. FOR every file in the dataset:
   2.1. Read file data and label
   2.2. Generate gaze cue matrix:
      2.2.1 FOR every frame within the cue window
      2.2.2 For all pairs of frames within the cue window
      2.2.3 Get gaze features and append to gaze cue matrix
   2.3. Generate pose cue matrix:
      2.3.1 FOR every frame within the cue window
      2.3.2 For all pairs of frames within the cue window
      2.3.3 Get pose features and append to pose cue matrix
   2.4. Generate AUs cue matrix:
      2.4.1 FOR every frame within the cue window
      2.4.2 For all pairs of frames within the cue window
      2.4.3 Get AUs features and append to AUs cue matrix
3. RETURN gaze cue matrix, pose cue matrix, AUs cue matrix

END

The spectrum of features values for gaze, pose and au's are different, hence the prediction model will not get any knowledge and the training process will fail. Generally, machine learning models expect the input have homogeneous values, which are values taken from the same space and describe the same phenomenon. On the other hand, features collected here describe three different spaces and the features values describe different meaning in each space collected (gaze, pose, and au's). Therefore, the tracking fille features were divided into three separated cues files.

$[g_0 \quad g_{1.....}g_{n-1}]^t$

$[p_0 \quad p_{1.....}p_{n-1}]^t$

$[Au_0 \quad Au_{1.....}Au_{n-1}]^t$

Where g_n, p_n, AU_n is a gaze, pose, and action units respectively ;

n = 0 to number of features per frame -1; and t is the time spent on each frame.

3.4 Modified GoogLeNet

The proposed model implements a modified version of the well-known and wide-used GoogleNet machine learning model to classify the emotions.

The key feature of GoogleNet is the inception module, which is designed to capture different levels of spatial features and perform dimensionality reduction using parallel convolutional layers with different filter sizes. The inception module enables the network to strike a balance between computational complexity and precision by averting the costly operation of processing all the data through a single convolutional layer.

In the original model, the input layer except images of multiple channels. Followed by an inception layer. The proposed model changes this single input model into three input layers. Each branch of the proposed architecture contains a stem, two inception layers and one max pooling layer before concatenating the three main branches, figure (3.4) illustrate the flow of the modified GoogleNet. The purpose of these layers is to get initial knowledge from the extracted cues, when the network reaches the layer before concatenation the values now are insights gathered by the model rather than values represent the original features (i.e., abstractions). After this step the structure continues as the original network.

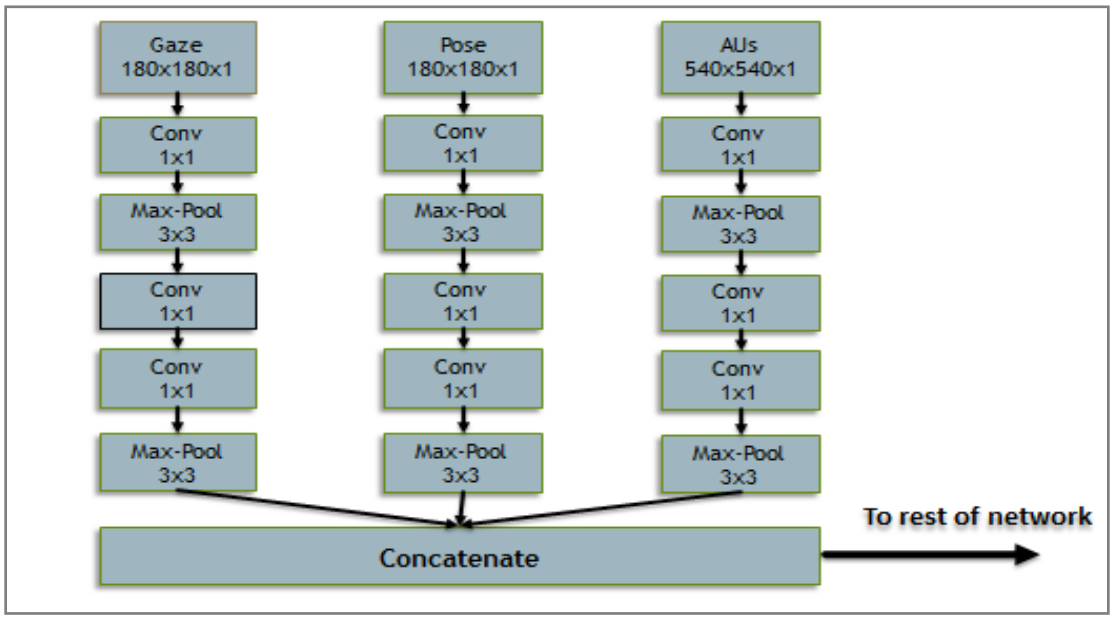


Figure 3.4:Modified GoogleNet architecture

Algorithm 3.3: Inception Block
INPUT: input_layer, number of filters
OUTPUT: output layer
BEGIN
1. Initialize the input layer and the number of filters for each path
2. Create the first path by applying Conv2D with the specified number of filters and activation function
3. Create the second path by applying Conv2D with the specified number of filters and activation function
4. Create the third path by applying Conv2D with the specified number of filters and activation function
5. Create the fourth path by applying Conv2D with the specified number of filters and activation function
6. Concatenate the outputs from all four paths to create the output layer
7. Return the output layer.
END

Algorithm 3.4: EmotioNet function:
INPUT: gaze features, pose features, and AUs features
OUTPUT: the classification result
BEGIN
1. Create input layers for gaze features, pose features, and AUs features.
2. Create a gaze subnet by applying the following layers:
   2.1. Convolutional layer with 64 filters, kernel size of (7,7), and ReLU activation function.
   2.2. Max pooling layer with pool size of (3,3).
   2.3. Convolutional layer with 64 filters, kernel size of (1,1), and ReLU activation function
   2.4. Convolutional layer with 192 filters, kernel size of (3,3), and ReLU activation.
   2.5. Max pooling layer with pool size of (3,3).
3. Create a pose subnet by applying the following layers:
   3.1. Convolutional layer with 64 filters, kernel size of (7,7), and ReLU activation function.
   3.2. Max pooling layer with pool size of (3,3).
   3.3. Convolutional layer with 64 filters, kernel size of (1,1), and ReLU activation function
   3.4. Convolutional layer with 192 filters, kernel size of (3,3), and ReLU activation.
   3.5. Max pooling layer with pool size of (3,3).
4. Create an AUs subnet by applying the following layers:
   4.1. Convolutional layer with 64 filters, kernel size of (7,7), and ReLU activation function.
   4.2. Max pooling layer with pool size of (3,3).
   4.3. Convolutional layer with 64 filters, kernel size of (1,1), and ReLU activation function
   4.4. Convolutional layer with 192 filters, kernel size of (3,3), and ReLU activation.
   4.5. Max pooling layer with pool size of (3,3).
5. Concatenate the results from steps 2, 3, and 4.
6. Apply nine inception blocks:
   6.1. Apply first Inception block (Algorithm 3.4) to the results from step 5.
   6.2. Apply second Inception block (Algorithm 3.4) to the results from step 6.
   6.3. Apply a max pooling layer to the output with pool size of (3,3).
   6.4. Apply a third Inception block to the output.
   6.5. Apply a fourth Inception block to the output
   6.6. Apply a fifth Inception block to the output
   6.7. Apply a sixth Inception block to the output
   6.8. Apply a seventh Inception block to the output
   6.9. Apply a max pooling layer with pool size of (3,3) the output

6.10. Apply a eighths Inception block to the output
6.11 Apply a ninth Inception block to the output
6.12. Apply a global average pooling layer for the resulted model network
7. Enforce a weights dropout layer with dropout percentage of 40%.
8.Generate the final classification dense layer with 5 classes (emotions classes) using SoftMax as an activation function.
9. RETURN the final model
END

3.5 Training Phase

The approach you described involves fitting two separate models, one for AS and another for TD. These models are trained on separate datasets, presumably containing features specific to each group. The purpose of this method is to evaluate the models' capacity to generalize their learning skills to new sets of features. By training separate models on AS and TD datasets, you can assess how well each model performs on their respective target groups. This helps evaluate whether the models can effectively capture the patterns and characteristics unique to each group.

Overall, this approach allows for a comprehensive evaluation of the models' generalization capabilities and their performance on new scopes of features, encompassing both AS and TD individuals.

During the training phase, the cues flow in batches to the model and traverse through the network structure.

The output of a single pass of the network is validated using the validation set to get an intermediate metrics these metrics are compared with the previously collected metrics to decide if the model is training in the correct direction. In

addition, this process measure if there is any overfitting on the training set. The process (form data-batch enters the model till the model validation is called an Epoch).

After completing all the training-validation epochs the model is now can be called a trained model (not the final model).

3.6 Test Phase

After the model has been trained and validated, its performance on the test set is assessed. This will provide an objective estimation of its generality. The generality of a model is its capacity to perform well with untrained and unseen data Evaluating the model generality depending on the nature of the task, so several metrics were be used to measure the generality of a model, including accuracy, precision, and recall.

Measuring the generality of a DL model is a crucial aspect of evaluating its performance and ensuring that it will perform well on new, unseen data. Using the evaluation phase, we can ensure that the model is accurate and generalizable, which is essential for applications in the real world.

3.7 summary

The proposed system design involves the development of a supervised deep learning model that is trained on a dataset containing eye-gaze, head-pose, and face action units tracking features of typical development people and autistic people. The model is intended to classify emotions based on the generated emotion cues extracted from the tracking files. The dataset is processed to obtain emotion cues, which are then divided into three separate cues files due to their different values and meanings. The model is built using a modified version of the GoogleNet deep

learning model, which is designed to capture different levels of spatial features and perform dimensionality reduction using parallel convolutional layers with different filter sizes. . The final model is evaluated on unseen data to assess its level of generality and ability to work with real data.

Chapter Four

Results and Discussion

4.1 Introduction

This chapter shows the significant results acquired from different experiments shown throughout the previous chapter. followed by a thorough comparison and evaluation of the model's outcomes. The chapter illustrates the differences between AS and TD individuals when the model was trained on separate groups of datasets. Finally, the generality evaluation of the model was conducted against un-seen data.

4.2 Data pre-processing

The OpenFace produces results data filtered from outliers and missing data. the only thing that needs to be cleaned in the resulted tracking data is the validity of the processed frame. The validation indicator in the features data file is the "success" filed which is a binary filed ("1" means frame is captured and processed correctly, and "0" otherwise). Frames with success value equal to "0" is eliminated from the training dataset. Table (4.1), illustrates the number of valid and invalid frames in the dataset.

Table 4.1:Counts for Valid and Invalid Frames

	Fear	Disgust	Joy	Anticipate	Sadness	Total
Invalid Frames	228	548	168	685	629	**2258**
	1.68%	2.88%	0.64%	0.65%	1.49%	**1.09%**
Valid Frames	13345	18485	26044	104671	41617	**204162**
	98.32%	97.12%	99.36%	99.35%	98.51%	**98.91%**
Total Frames	13573	19033	26212	105356	42246	**206420**

The table above indicates that the total frames lost is 1% from the total number of frames captured which indicates the high accuracy of the OpenFace API.

4.3 Results of the Implementation of the Proposed System

The proportion employed to divide the dataset relies on its size and the task's characteristics. A prevalent ratio for dividing the data into training and testing sets is 80/20 or 70/30, with 80% or 70% of the data allocated for training and the remaining 20% or 30% designated for testing. When dealing with larger datasets, a split of 90/10 or 95/5 may be sufficient. Occasionally, a three-way split is utilized, which involves creating a validation set to fine-tune hyperparameters and prevent overfitting. The ratio for the three-way split can be 60/20/20 or 70/15/15, depending on the dataset's size.

Overall, dataset splitting is an essential step in machine learning, and the ratio used in splitting the dataset depends on the size and nature of the task.

The dataset was split into different ratios (60% training, 40% validation), (70% training, 30% validation) and (80% training, 20% validation)

The experiments performance showed that the ratio of 70/15/15 (70% training, 15% validation, and 15% test) achieved significant results. therefore, this ratio was implemented in this model. The test set is primarily used to test the model generality against unseen data during the training phase. Table(4.2) below states the influence of split ratio on model accuracy and error rate.

Table 4. 2: Dataset splitting ratio effect on Accuracy and Loss.

SPLIT RATIO(TRAIN/VAL)	**ACCURACY%**	**ERROR RATE%**
60%-40%	82.97%	0.17
70%-30%	**87.12%**	**0.12**
80%-20%	82.85%	0.17

GoogleNet uses stochastic gradient descent (SGD) to optimize its model parameters during training. However, GoogleNet incorporates a modified version of SGD known as "batch normalization." This technique, illustrated by the equation below, aids in speeding up the training process and enhancing the model's stability by normalizing the activations within each mini-batch.

$$\boldsymbol{x_{k+1} = x_k - \eta_k \left(\frac{1}{|B_k|} \sum_{i \in B_k} \nabla f_i(x_k) \right)} \qquad (4.1)\ [67]$$

Where x_k is the weights at step k, η_k is the learning step size at k, $|B_k|$ is the batch size, f_i is the objective function, and $\nabla f_i(x_k)$ is the gradient function.

Large batch sizes have been found to result in decreased accuracy and limited prediction generalization, as noted by [67][68]. In order to determine the appropriate batch size for the proposed model, a series of experiments were conducted using different batch sizes, namely 32, 64, 126, 256, and 512. The performance of the model on both the training and testing datasets revealed two distinct trends, as depicted in Table 4.3.

Table 4.2: Batch size experiments results

Batch Size	**Training Accuracy (A_{train})**	**Validation Accuracy ($A_{validate}$)**	**A_{train}- $A_{validate}$**
32	95.91%	90.68%	5.23
64	**97.76%**	**96.95%**	**0.81**
128	98.69%	86.32%	12.37
256	99.07%	82.47%	16.6
512	99.49%	83.16%	16.33

Larger batch sizes resulted in higher training accuracy, but this led to underfitting issues (99.49% training accuracy compared to 83.16% validation accuracy with a batch size of 512). Determining the optimal batch size is a task-specific procedure, and the best trade-off between training and validation accuracy, with minimal discrepancy, was observed with a batch size of 64.

Table 4.4: The Final Model Parameters are Stated in the Table Below

Hyper-parameter	**Value**
Dataset splits ratio	70% Training, 30%Validation/Test
Mini-Batch Size	64
Weights initialization	Activation initialization algorithm ReLu orthogonal Softmax Glorot Uniform
Loss Function	Categorical Cross-entropy
Optimizer Learning rate	Adam [69] 0.001

4.4 Results of the Proposed emotional Model

In this section, the outcomes of the model's learning phase will be presented. As described in Section 3.5 of Chapter 3, about the training and validation processes of the model. So, the accuracy of the model training are depicted in Figure (4.1). The training accuracy serves as an indicator of the model's precision in predicting all classes. The figure below states that the model accuracy nearly 99% and 98% for training and validation respectivly.

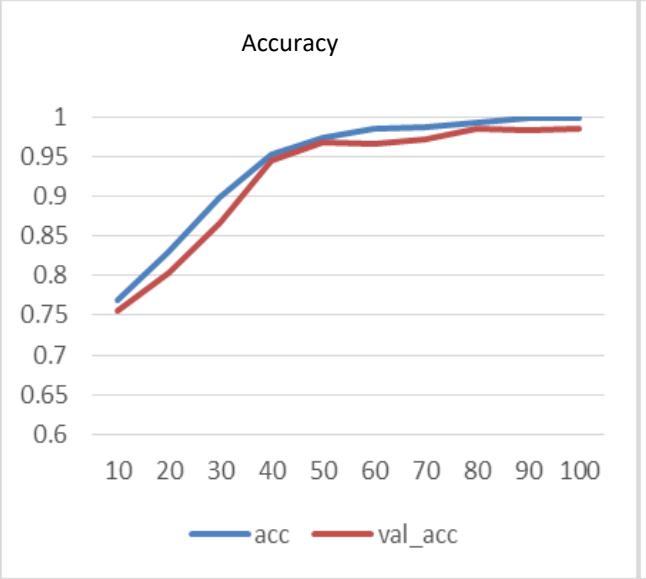


Figure 4.1: Accuracy of model training and validation**.**

4.5 Evaluating the Model

The model, as mentioned in Chapter 3, was trained and validated on data for individuals with AS and TD. The purpose of this procedure was to measure the model performance, as well as The performance of the model was evaluated using AS and TD data to investigate the significant variation in characteristics that discriminate between those two groups.

Hence, the results discussed in this section concern the emotional model trained on AS dataset only. The accuracy scores were 99.88% and 98.54% for model training and validation respectively. The accuracy was calculated statistically to assess the total correct predictions of the model (i.e. the total number of all correct prediction cues for all five classes). Formula 2.16 states the formula of the model's accuracy calculation.

As accuracy above used as a metric to measure the average all of the model performance based on all class prediction. To evaluate model performance based

on true class prediction which match the ground truth data , a Confusion matrix was used.

The model has a high level of performance among all classes in both True (predicting the right class) and False (saying that this is not the required class) classifications. . For better understanding of the model performance for inter-class prediction, the model was tested on an un-seen dataset. The model test-dataset was excluded from the training dataset before beginning the training phase of the model and set aside as the un-seen dataset. The benefit of the un-seen data was to challenge the model architecture to produce the correct emotional state. The correct classification means having the correct label for the given cue, for the un-seen data was to challenge the model architecture to produce the correct affective. The model was tested on 22953 labeled cues accommodating all the five emotional states. A correct (Positive) prediction should match the ground truth label, or it would be incorrect (Negative) prediction.

The statistical results of the model's test experiments are summarized in a confusion matrix Figure (4.2), it states the correct and incorrect prediction rates for each emotional class. The rows, in confusion matrices, describe the model's class prediction, while the columns are the true classes obtained from the labelled dataset. The highest classification rate was (99%) scored for both disgust and Anticipation classes, while the classification rate for sadness was (98%) and fear was (93%). The lowest prediction score was (92%) for the joy class. There were minor variations in prediction rates between all classes and that can be realized from the monotone of the classes cells' color. These results indicate a robust performance of the model and high generality skills of the emotional computational model towards un-seen data. In addition to the high recognition

rates for each class in the confusion matrix, other significant metrics can determine the model generality.

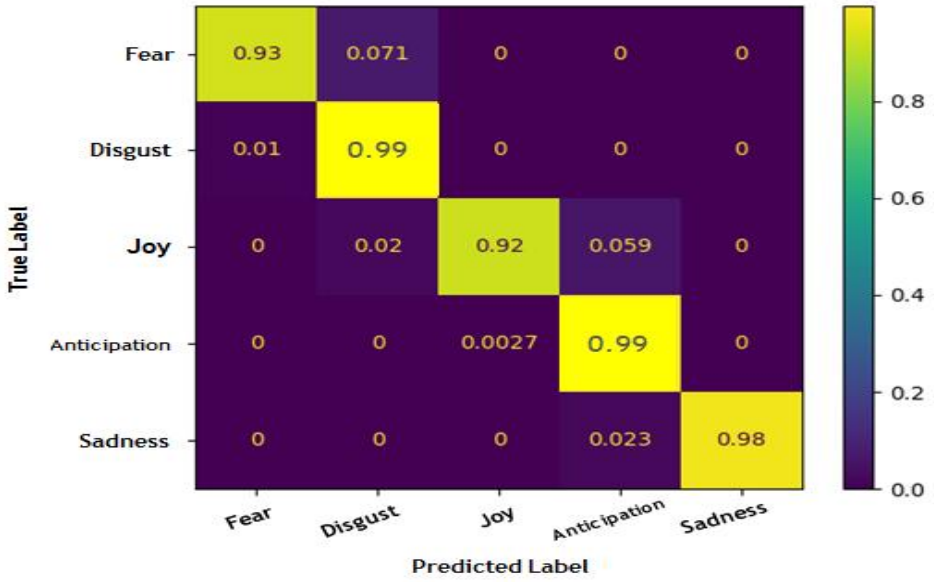


Figure 4.2: Confusion matrix for the model tested on data for People with AS.

For further optimized evaluation of the model performance, other indicators were used to assess the model's performance as an emotional classifier .The statistics used were recall (0.96), precisions (0.97).

The above results in this section show the performance of the developed model when it is trained on data from people with AS and tested on unseen data. The same procedure stands when training the model on data from people with TD and testing on unseen data from the same group, and this will be discussed in the next part.

The model was also tested on data from a person with TD on model trained on data from the same group. The confusion matrix in Figure 4.3 shows significant results yield from this test.

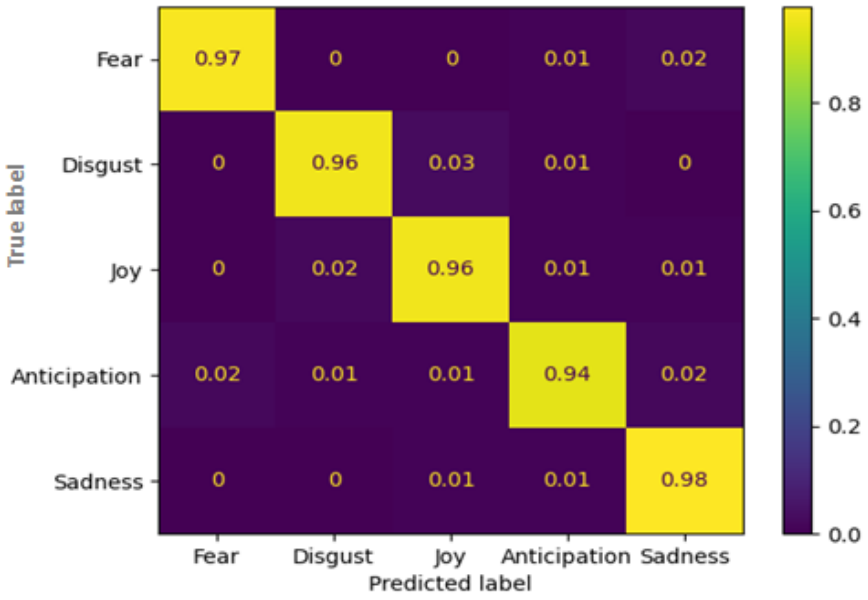


Figure 4.3: Confusion matrix for the model tested on data for People with TD.

The results drawn from the confusion matrix in (Figure 4.3) shows that the classes with the highest TP ratio were the sadness with accuracy rate of 98% and the fear emotional state with accuracy rate of 97%.

While disgusted and joy have the same TP rate of 96%, the lowest TP rate was 94% for the anticipation class.

Also here, other indicators were used to evaluate the model's performance as an emotional classifier, namely Precision (Equation 2.17) and Recall (Equation 2.18). The results of the statistics were (0.96) and (0.96), respectively

4.6 Features Analyzing

The proposed emotional model achieved significant results in modeling individuals' emotional behaviors. The use of deep learning techniques, specifically GoogleNet, has likely played a crucial role in achieving these efficient results. GoogleNet is a deep convolutional neural network (CNN) that is known for its ability to extract meaningful features and patterns from various input cues and handle complex visual recognition tasks effectively. In Addition, the success of emotions recognition and classification may also be attributed to the way in which cues are generated and twinned with GoogleNet's structure. The process of building and generating cues to produce emotions is likely designed to capture diverse aspects of human emotional expression, allowing the model to better understand and predict emotional states accurately. twinning or integrating three input channels (facial expressions, eye gaze, and head movements) with the GoogleNet structure, increased the network's capacity to learn complex relationships between input cues and emotions.

The three feature-sets (AUs, eye-gaze, and head-pose) are considered as the main source channel to extract the emotions in this study. (Figures 4.4, and 4.5) illustrate the contribution ratio of each features set per emotional state. The ratios were calculated by counting the probability contribution of the significant values for each feature (values that cause variation ns in the cues sequence or features with values greater than 0).The results presented in Figure 4.4 describe the probability of feature-groups contribution ratios for the model trained on data from individuals with AS.

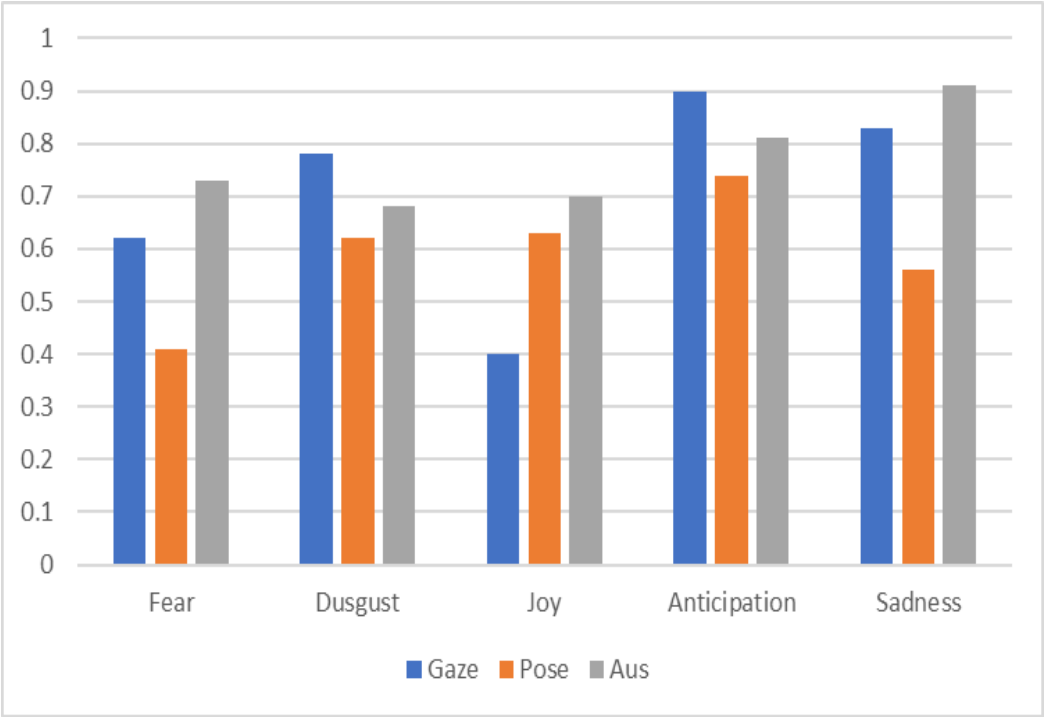


Figure 4.4: The results of features contribution for each emotional class of AS.

The results show that there is little variation in using three scopes of feature-sets by AS to express their emotional states. the dominant features-set was AUs with highest ratio for sadness state 0.9% and lowest ratio was in the disgust state 0.67%. The overall use of eye-gaze and head-pose was roughly equal which is an indication that AS use more intense movements in their eyes and head to express their emotional state. One of the features that can be derived from the above figure is when people with AS experience sadness state they tend to use their facial expressions more than eye or head movements (means a relatively fixed eye-gaze with little head movements). The highest use of eye-gaze was in two states of sadness, and anticipation with 0.9%, and 0.78% respectively which combined with high frequencies of head movements. The above ratios can provide evidence that models concerned with AS should take all the facial expressions in addition to eye-gaze and head-pose to assist the prediction model learning skills.

(Figure 4.5) shows the results of the feature-sets contribution ratio for the model trained with data from individuals with TD. It also shows the vast difference in using communication channels comparing with individuals with AS (Figure 4.4). The model learned almost all its skills from facial expressions (i.e., action units). This comes from the relatively high experience of TD individuals in using face gestures to express their emotional states with less demand on eye-gaze and head-pose. The figure exhibits that the model can learn about 92% of its prediction skills from AUs alone and that for a model not dedicated to use AUs alone. All these conclusions draw that emotional models dedicated for TD can use AUs alone to get high accuracies of predictions.

(Figures 4.5 and 4.6) presented the participation ratio for each features-set along with all classes for each dataset group. The two figures displayed the main differences in displaying emotional states for individuals with AS, and TD.

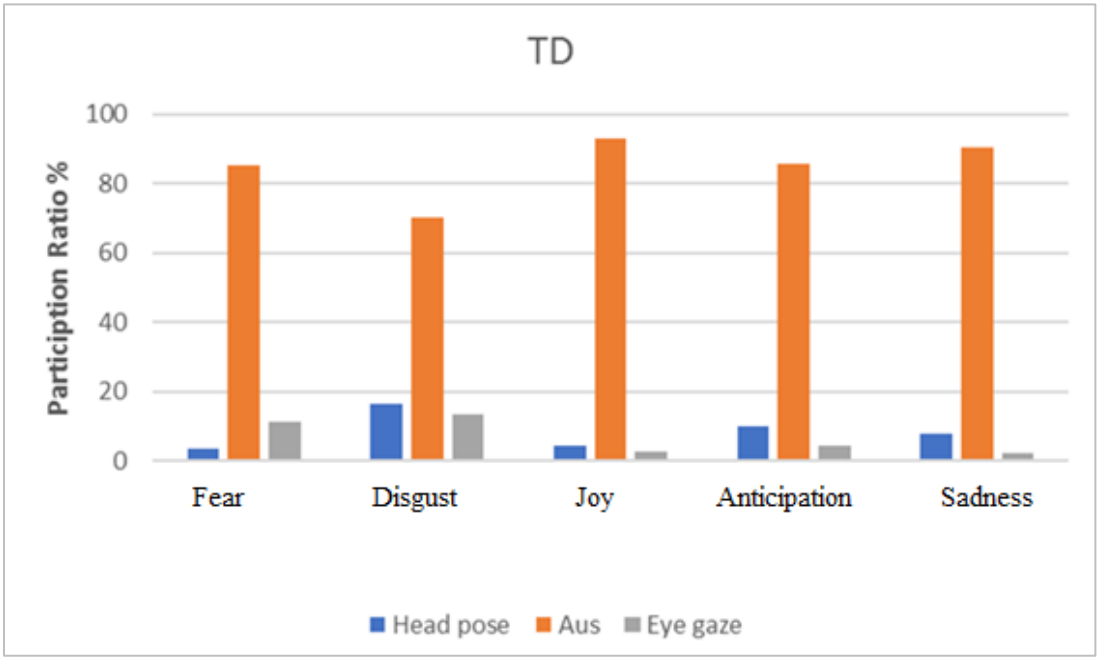


Figure 4.5: The results of features contribution for each emotional class of TD .

4.6.1 Facial Action Units

Table 4.5 (visualised in Figures 4.6 and 4.7) shows the probability of frequencies in each class of emotions for each dataset group (AS and TD). The probabilities ratio in the table below refer to occurrence of single action units in each cue along all video-sequences.

Table 4.5: Probabilities of AUS frequencies between AS and TD for all class's

	Fear		**Disgust**		**Joy**		**Anticipation**		**Sadness**	
	ASD	TD	ASD	TD	ASD	TD	ASD	TD	ASD	TD
AU01	0.0065	0.0081	0.018	0.0056	0.0147	0.0134	0.0488	0.0435	0.0215	0.0675
AU02	0.0088	0.0103	0.0177	0.0065	0.0189	0.0156	0.0578	0.0649	0.036	0.0851
AU04	0.0107	0.0291	0.0299	0.0219	0.0079	0.0196	0.0538	0.1272	0.04	0.1092
AU05	0.0404	0.0386	0.0515	0.0173	0.0619	0.0124	0.2907	0.2625	0.1265	0.1563
AU06	0.0149	0.0176	0.0112	0.0043	0.0567	0.0598	0.0379	0.0672	0.0326	0.0396
AU07	0.0216	0.0567	0.0392	0.0197	0.0604	0.0622	0.1327	0.3076	0.0551	0.1579
AU09	0.0031	0.0019	0.0055	0.002	0.0035	0.0027	0.0145	0.0115	0.0095	0.0441
AU10	0.0202	0.0444	0.0193	0.0095	0.0682	0.0662	0.0891	0.1668	0.0417	0.1355
AU12	0.0169	0.0196	0.0091	0.0036	0.0712	0.062	0.0282	0.0669	0.0221	0.0388
AU14	0.0375	0.0564	0.0276	0.0282	0.0889	0.0684	0.2396	0.3473	0.0893	0.2196
AU15	0.0069	0.0058	0.0191	0.004	0.0192	0.02	0.0525	0.0408	0.027	0.0444
AU17	0.0184	0.0185	0.0334	0.0068	0.0358	0.0197	0.1051	0.0904	0.0633	0.0836
AU20	0.0043	0.0058	0.0134	0.0028	0.0084	0.0108	0.04	0.0311	0.021	0.0515
AU23	0.0061	0.0321	0.0079	0.0111	0.0156	0.035	0.024	0.2355	0.018	0.0876
AU25	0.0146	0.0182	0.0225	0.0072	0.0318	0.023	0.0676	0.0975	0.0428	0.1155
AU26	0.0118	0.011	0.0175	0.0046	0.0217	0.0134	0.0451	0.0617	0.0327	0.0727
AU28	0.0001	0.0001	0.0009	0.0003	0.0006	0.001	0.002	0.0039	0.0039	0.002
AU45	0.0136	0.0143	0.025	0.0085	0.0298	0.0173	0.1005	0.0934	0.0522	0.1092

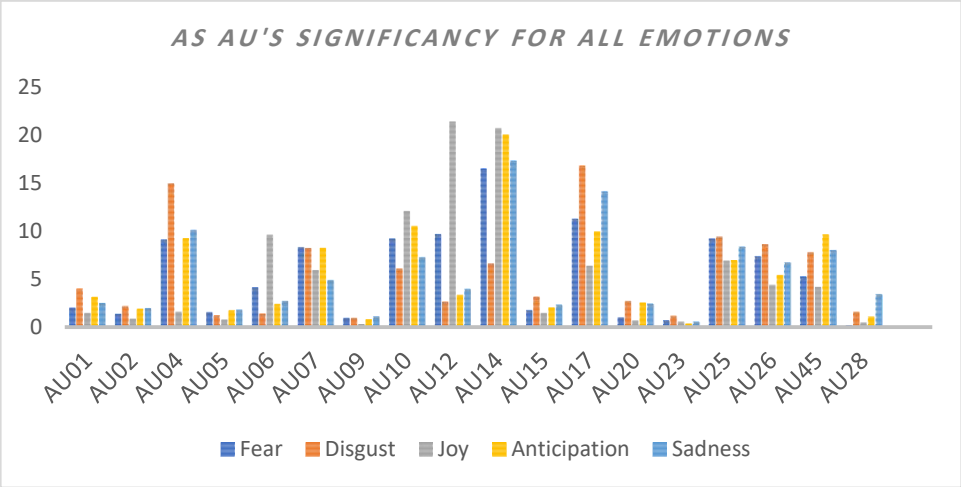


Figure 4.6: Bar chart presents the highest frequency differences of AUS in each class for AS

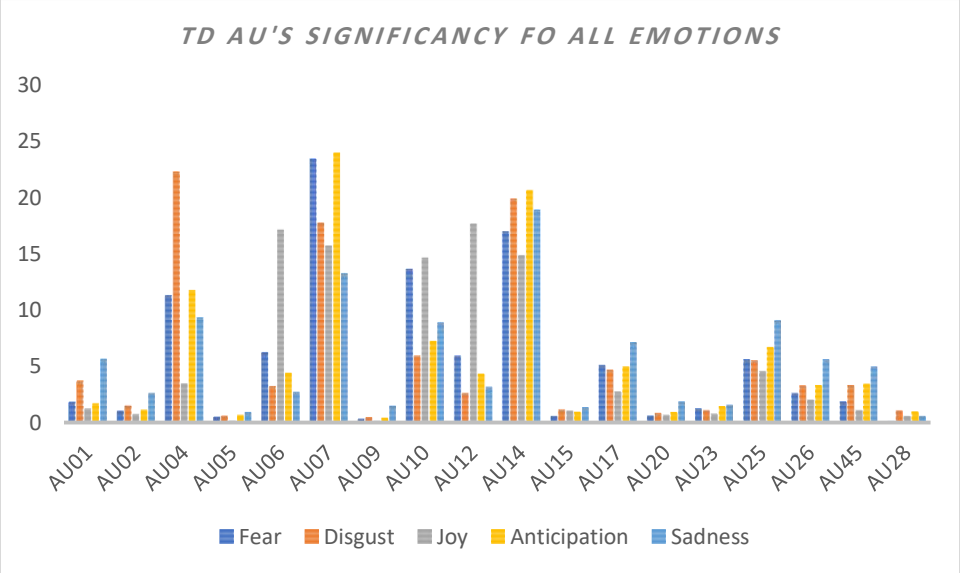


Figure 4.7: Bar chart presents the highest frequency differences of AUS in each class for TD

(Figures 4.6 and 4.7) visualize a bar graph to demonstrate the differences in an emotional state for AUs as extracted from the predictions of the affective model between individuals with AS and TD.it can be see that the same AU's were expressed for the same emotions, but they did not have the same intensity.

Generally, the analysis of facial action units shows persistent differences between AS and TD when experiencing emotional states.

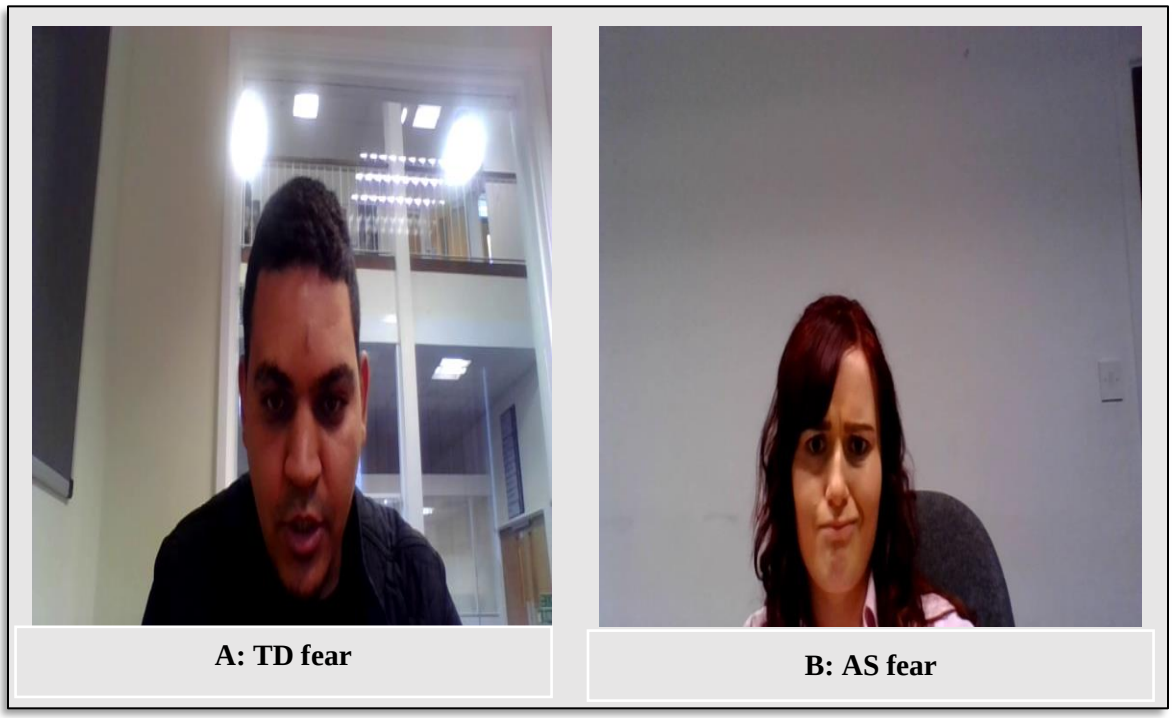
For example, In (Figure 4.6), the facial action units (AUs) that had the highest likelihood of indicating fear were AUs 4, 10, 12, 14, and 17 for individuals with Autism Spectrum (AS). On the other hand, in (Figure 4.7), among the typically developing (TD) group, the highest intensity was observed for of AU7, along with AU's 4, 6, 10, and 12. Additionally, it was noticeable that the intensity of AU17 decreased across the TD group when expressing the same emotion.

When taking disgust emotion, the highest values appeared significantly in AUS 4, 17 for AS, and AUs 4,7, and 14 for TD.

joy emotion has the highest values of AUS in 6,7,10,12 and 14 (as illustrated in Figure 4.7) for TD . While in the AS group (as in Figure 4.6), it was observed that the intensity of the action units of the upper face decreased, while the density of the action units of the lower face increased, compared to the group of TD.

in Anticipation emotion, the probability of AU 14 occurrence in AS was higher than TD, in contrast, AU 7 occurrence was higher in TD than AS.

In general, individuals with TD tend to have more extensive and intense Action Units (AUs) compared to individuals with AS. Additionally, individuals with AS may exhibit different facial expressions compared to neurotypical individuals (TD). For example, (Figure 4.8) illustrates a sample from the dataset featuring a typical person and an autistic person expressing fear. The autistic person displayed AU6 (lip corner puller) and AU25 (lips parting), which typically indicate happiness according to Ekman's classification mentioned in (section 2.3.1).



A: TD fear

B: AS fear

Figure 4.8(A) frame of fear emotion for TD, (B) frame of fear emotion for AS

For better clarification, a thesis was undertaken to compare scientist Ekman's discoveries on detecting fear in individuals with typical development (TD) through facial cues like raised eyebrows, widened eyes, and a tense mouth (illustrated in Figure 4.9) with the results obtained from the developed model (depicted in Figure 4.10), which relies on action units. The analysis leads to the conclusion that Ekman's observations support the findings obtained from the proposed emotional model. However, the contribution of this thesis and the model acknowledges that individuals with autism spectrum (AS) exhibit variations in facial expressions. The comparison reveals that individuals with AS may display atypical manifestations of emotional state, emphasizing the importance of considering context and individual differences.

Basic expressions	Involved Action Units
Surprise	AU 1, 2, 5,15,16, 20, 26
Fear	AU 1, 2, 4, 5,15,20, 26
Disgust	AU 2, 4, 9, 15, 17
Anger	AU 2, 4, 7, 9,10, 20, 26
Happiness	AU 1, 6,12,14
Sadness	AU 1, 4,15, 23

Figure (4.9) AUs in fear emotion [70]

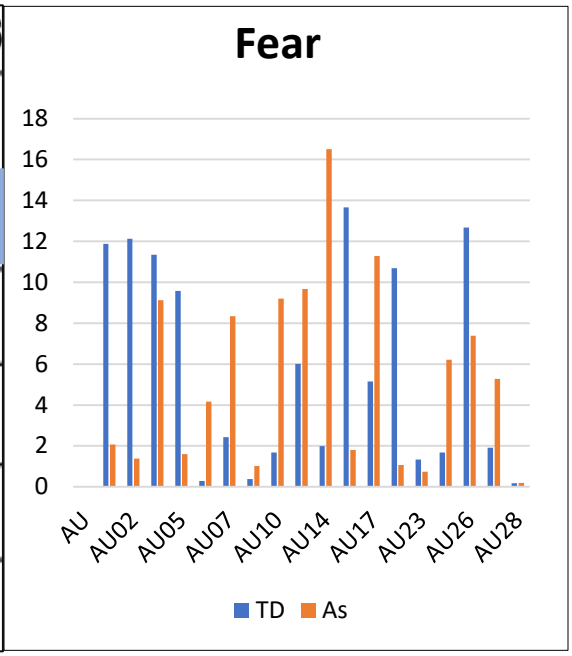


Figure (4.10) difference of AUs in fear emotion between AS and TD

In (Figure 4.11), a framework is presented to understand the experience of joy based on the dataset. The framework is supported by a bar graph (also labeled 4.12), which illustrates the findings. Upon analyzing the graph, it is observed that the facial expressions associated with joy in the average person are more intense compared to those of individuals with autism. Both the average person and the person with autism exhibit similar patterns in expressing feelings of joy, but with variations in frequency and intensity.

Figure 4.11(A) frame of joy emotion for TD, (B) frame of joy emotion for AS

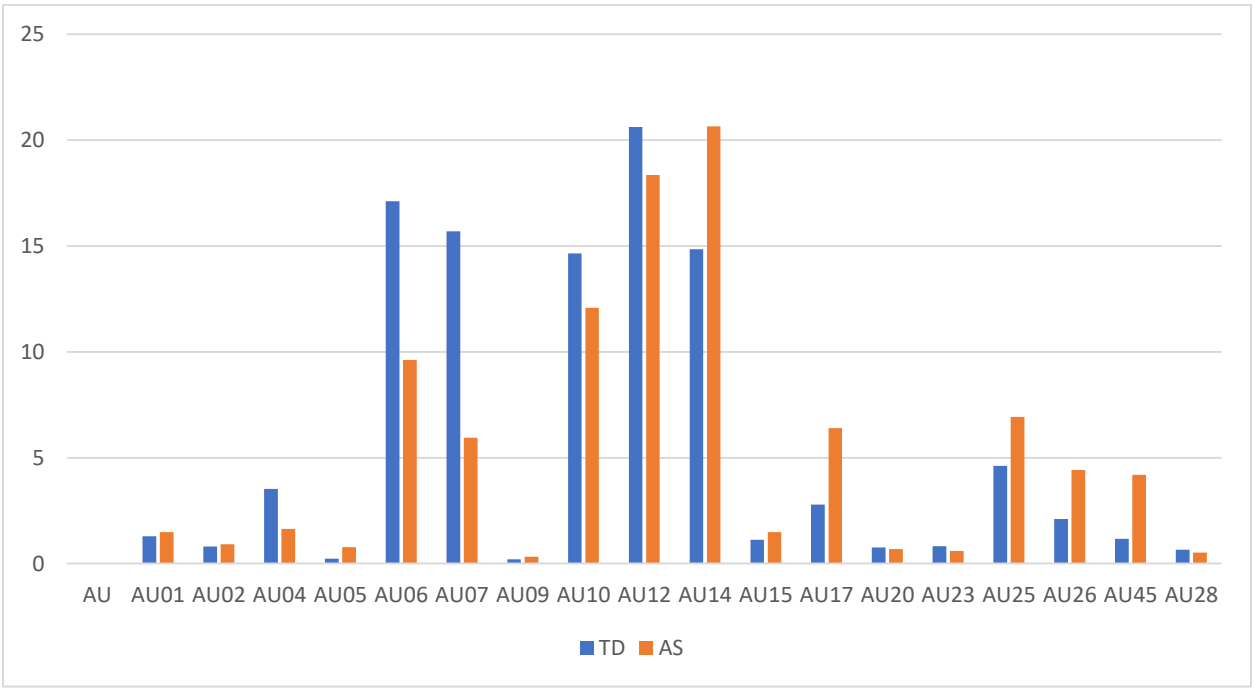


Figure (4.12) difference of AUs in joy emotion between AS and TD

Also, different investigations used by researchers have proved that there are differences in expressing emotions between individuals with AS and TD [24][25]. In addition to these claims, to further ensure the validity of our assumptions, another test was conducted. This testing involved crossing data of individuals with Autism Spectrum (AS) with a model trained on individuals with typical development (TD) data, and vice versa. The purpose of this test was to demonstrate the validation of variations in facial features between these two groups in expressing the same emotions. The results can be shown in below confusion matrix (Fig 4.13).

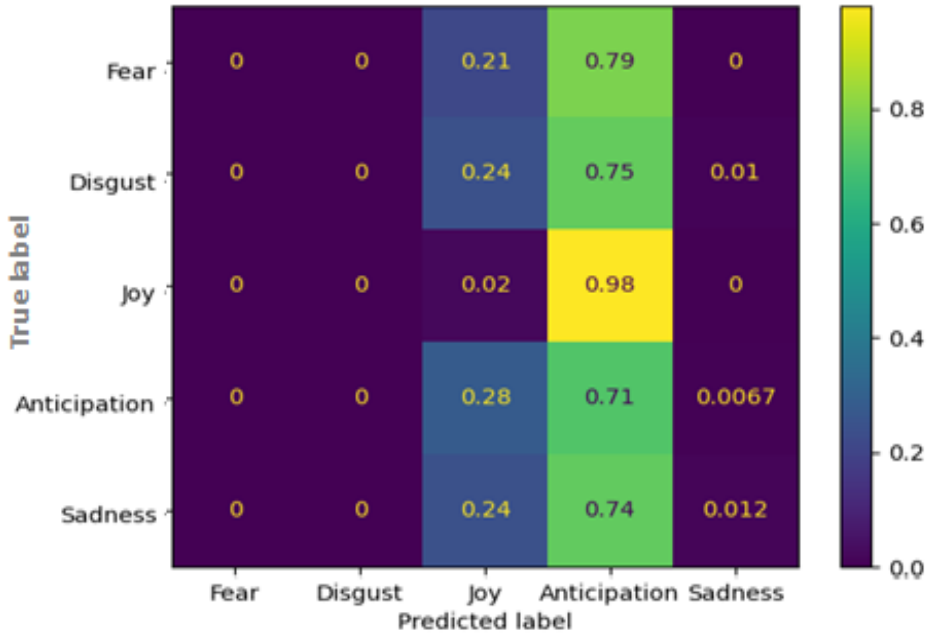


Figure 4.13: Confusion matrix of the emotional model trained on AS data and tested on TD data

The results from the Confusion matrix indicated a decrease in the accuracy of the emotional model when tested with TD data. This decrease in accuracy suggests that there are notable differences in facial features between the AS and TD populations. Consequently, the emotional model trained on AS data might not be capable of accurately interpreting the emotional states of TD individuals because of these feature differences.

4.7 Model Generality

The model's performance was validated using unseen data, demonstrating its efficiency when applied to these data, as mentioned earlier. Obtaining public available datasets for people with autism posed difficulties due to ethical considerations and the need for consent agreements. Consequently, the proposed work utilized a dataset specifically created for scientific research, with participants agreeing to the use and publication of their images and videos.

However, some student who took part in the experiments related to this dataset did not grant consent to publish their images. Nonetheless, their images captured a completely different environment during recording, presenting an excellent opportunity to utilize these images as data for testing the model's generality. The proposed model underwent evaluation using these images and demonstrated its ability to generalize to another dataset with a different environment, achieving an accuracy rate of about 90%. This high accuracy ratio attests to the model's effectiveness when dealing with unseen data.

4.8 Model environment

In real-time applications, particularly in Human-Computer Interaction (HCI), the speed of a model is a crucial factor. Model speed refers to the time taken from receiving an input to producing an output or response. To evaluate the proposed model's speed, tests were conducted on a computer with GPU NVADIA GeForce GTX 1050 2GB, using an Intel® Core™ i7-7300HQ CPU @2.50 GHz processor, 16 GB of RAM, and a 64-bit Windows 10 operating system.

4.9 Model Comparing

The proposed study produced an emotional model for individuals with AS using spontaneous natural data captured using only a webcam without the use of invasive tools. The model showed significant results, with an average accuracy of 97.54%. In most cases, the results of this study were generally consistent with other studies on the existence of differences in emotional expression between people with AS and those with TD. Most studies investigated facial expressions of individuals with AS were built to serve psychological studies, therapists, treatment, or even educational skills. Previous studies utilized two categories of data: natural spontaneous data obtained in unsupervised environment, and induced data collected by employing stimuli to provoke emotions or identify facial expressions within a controlled environment, resulting in artificial behaviors. Consequently, the comparison in this section was based on the technical methods followed to produce the emotional model, and tools that capture the features. also, The axis for the comparison As shown in the table (4.6) was based on comparing the results from models which used deep learning against other machine learning techniques. The dynamic complexity of features in natural-spontaneous data can be further

processed using advanced methods like dimensionality reduction or computer-vision and image-processing.

Table (4.6) A comparison of proposed system with related work

Reference	**Type of dataset**	**Tools that used to capture features**	**Methodology**	**Input channels**	**Acc.**
[21]	Induced Data	1.camera 2. robot 3. sensor	Support Vector Machine classifier	-Action unit -Head movement	93.6%
[22]	Induced Data	1.webcam 2.EPOC+headse	-CNN -LSTM	-facial expressions -EEG signals	-99.81% -87.25%
[23]	Spontaneous Data	Camera	-Viola jones algorithm -adaboost	-AUs	85.97%
[1]	Induced Data	Webcam	CNN(deep ResNet)	facial expressions	75%
[5]	Spontaneous Data	Webcam	-CNN -LSTM	-facial expressions -head movement -eye gaze	90.06%
Proposed	Spontaneous Data	Webcam	CNN(GoogLeNeet)	-facial expressions -head movement -eye gaze	97.54%

The research in [5] is closely similar to the proposed system in terms of the data set, input channels, and tools utilized to capture features. However, there is a distinction in extracting complex emotions, as it encompasses a wider range of expressions. The researcher employed deep learning algorithms such as CNN and

LSTM, which yielded an average accuracy of 90.06%. In comparison, the proposed system achieved a higher accuracy rate of 97.54% by using Googlenet and the methods which used to generate Cues vector.

4.10 Conclusion

This chapter presented the testing and evaluation results for the proposed model. The model evaluation was tested and evaluated based on unseen data. the purpose for testing and evaluating to prove the power of model generality toward new data. Also, it stated the power of googlenet towards features extraction and prediction. Moreover, the differences in features expressions between two groups AS and TD were extracted. The features results showed that there are significant variations in features for the same emotions which expressed by two groups. Comparing models result with other previous works clarify the model efficiency by accuracy about 97.54%.

Chapter Five

Conclusion and Future Work

5.1 Conclusions

The following are the main conclusions gained from the results obtained using the proposed system to detect the emotional state of people with Asperger's syndrome:

1- Natural ,spontaneous data improved emotional state recognition without costly sensors in controlled environments.

2- Deep learning techniques outperformed shallow systems by extracting abstract features, incorporating facial expressions, head movements, and eye gaze to predict behaviors in individuals with AS, achieving high accuracy and generality for real-time applications,

3- The proposed model achieved high accuracy in inferring basic emotions, with a 97.54% generality towards unseen data for individuals with AS. The model produces reliable results in standard time, making it suitable for real-time applications and facilitating interaction and communication between people.

4- The accuracy of the emotional model, which was trained on AS data, dropped by approximately 27% when tested on TD data. This highlights the importance of recognizing variations in emotional expression between individuals with (AS) and (TD) individuals, confirming the distinctions in features established in this thesis.

5.2 Future works

The following is recommended for future works:

1-Voice Analysis: Develop technology for subtle tone, pitch, and rhythm changes in voice to enhance emotional inference in ASD individuals, alongside visual cues.

2. In futures there is a need to involve other types of Autism, this will help increase model generality through increasing the size of the autistic part of the dataset.

3- Cultural Differences: Conduct cross-cultural studies to understand the relationship between autism, cultural norms, and emotional expression, fostering more inclusive interventions and comprehensive autism understanding. Emotion expression varies based on cultural norms and values

References

References

[1] Bah, I., & Xue, Y. (2022). Facial expression recognition using adapted residual based deep neural network. *Intelligence & Robotics*, *2*(1), 78-88.

[2] Hassan, A., Pinkwart, N., & Shafi, M. (2021). Serious games to improve social and emotional intelligence in children with autism. *Entertainment computing*, *38*, 100417.

[3] Kuo, S. S., & Eack, S. M. (2020). Meta-analysis of cognitive performance in neurodevelopmental disorders during adulthood: Comparisons between autism spectrum disorder and schizophrenia on the Wechsler Adult Intelligence Scales. *Frontiers in Psychiatry*, *11*, 187.

[4] Sharma, S. R., Gonda, X., & Tarazi, F. I. (2018). Autism spectrum disorder: classification, diagnosis and therapy. *Pharmacology & therapeutics*, *190*, 91-104.

[5] Dawood, A., Turner, S., & Perepa, P. (2018). Affective computational model to extract natural affective states of students with Asperger syndrome (AS) in computer-based learning environment. *IEEE Access*, *6*, 67026-67034.

[6] Faras, H., Al Ateeqi, N., & Tidmarsh, L. (2010). Autism spectrum disorders. *Annals of Saudi medicine*, *30*(4), 295-300.

[7] Danforth, A. L., Struble, C. M., Yazar-Klosinski, B., & Grob, C. S. (2016). MDMA-assisted therapy: a new treatment model for social anxiety in autistic adults. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *64*, 237-249.

[8] Joseph, L., Pramod, S., & Nair, L. S. (2017, December). Emotion recognition in a social robot for robot-assisted therapy to autistic treatment using deep learning. In *2017 International Conference on Technological Advancements in Power and Energy (TAP Energy)* (pp. 1-6). IEEE.

[9] Nikopoulou, V. A., Holeva, V., Kerasidou, M. D., Kechayas, P., Papadopoulou, M., Vrochidou, E., ... & Kaburlasos, V. G. (2020). Identifying linguistic cues; towards developing robots with empathy in autism interventions. *Journal of Clinical Medicine of Kazakhstan*, *2*(56), 27-33.

[10] Eddy, C. M. (2019). What do you have in mind? Measures to assess mental state reasoning in neuropsychiatric populations. *Frontiers in psychiatry*, *10*, 425.

[11] Dantas, A. C., & do Nascimento, M. Z. (2022). Recognition of emotions for people with autism: An approach to improve skills. *International Journal of Computer Games Technology*, *2022*, 1-21.

[12] Lisetti, C. L. (1998). Affective computing. *Pattern Analysis & Applications*, *1*(1), 71-73.

[13] Sivasangari, A., Ajitha, P., Rajkumar, I., & Poonguzhali, S. (2019). Emotion recognition system for autism disordered people. *Journal of Ambient Intelligence and Humanized Computing*, 1-7.

[14] Steiner, H., Keplinger, F., Schalko, J., Hortschitz, W., & Stifter, M. (2015). Highly efficient passive thermal micro-actuator. *Journal of Microelectromechanical Systems*, *24*(6), 1981-1988.

[15] Alqahtani, F., Katsigiannis, S., & Ramzan, N. (2020). Using wearable physiological sensors for affect-aware intelligent tutoring systems. *IEEE Sensors Journal*, *21*(3), 3366-3378.

[16] Ristea, N. C., Duţu, L. C., & Radoi, A. (2019, October). Emotion recognition system from speech and visual information based on convolutional neural networks. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)* (pp. 1-6). IEEE.

[17] O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., ... & Walsh, J. (2020). Deep learning vs. traditional computer vision. In *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1* (pp. 128-144). Springer International Publishing.

[18] Loddo, A., Loddo, M., & Di Ruberto, C. (2021). A novel deep learning based approach for seed image classification and retrieval. *Computers and Electronics in Agriculture*, *187*, 106269.

[19] Rajkomar, A., Lingam, S., Taylor, A. G., Blum, M., & Mongan, J. (2017). High-throughput classification of radiographs using deep convolutional neural networks. *Journal of digital imaging*, *30*, 95-101.

[20] Drimalla, H., Baskow, I., Behnia, B., Roepke, S., & Dziobek, I. (2021). Imitation and recognition of facial emotions in autism: a computer vision approach. *Molecular autism*, *12*, 1-15.

[21] Silva, V., Soares, F., Esteves, J. S., Santos, C. P., & Pereira, A. P. (2021). Fostering emotion recognition in children with autism spectrum disorder. *Multimodal Technologies and Interaction*, *5*(10), 57.

[22] Hassouneh, A., Mutawa, A. M., & Murugappan, M. (2020). Development of a real-time emotion recognition system using facial expressions and EEG based on machine learning and deep neural network methods. *Informatics in Medicine Unlocked*, *20*, 100372.

[23] Singh, A., & Dewan, S. (2020). AutisMitr: Emotion recognition assistive tool for Autistic Children. *Open Computer Science*, *10*(1), 259-269.

[24] Manfredonia, J., Bangerter, A., Manyakov, N. V., Ness, S., Lewin, D., Skalkin, A., ... & Pandina, G. (2019). Automatic recognition of posed facial expression of emotion in individuals with autism spectrum disorder. *Journal of autism and developmental disorders*, *49*, 279-293.

[24] Manfredonia, J., Bangerter, A., Manyakov, N. V., Ness, S., Lewin, D., Skalkin, A., ... & Pandina, G. (2019). Automatic recognition of posed facial expression of emotion in individuals with autism spectrum disorder. *Journal of autism and developmental disorders*, *49*, 279-293.

[25] Król, M. E., & Król, M. (2019). A novel machine learning analysis of eye-tracking data reveals suboptimal visual information extraction from facial stimuli in individuals with autism. *Neuropsychologia*, *129*, 397-406.

[26] Samad, M. D., Diawara, N., Bobzien, J. L., Taylor, C. M., Harrington, J. W., & Iftekharuddin, K. M. (2019). A pilot study to identify autism related traits in spontaneous facial actions using computer vision. *Research in Autism Spectrum Disorders*, *65*, 14-24.

[27] Kleinginna Jr, P. R., & Kleinginna, A. M. (1981). A categorized list of motivation definitions, with a suggestion for a consensual definition. *Motivation and emotion*, *5*(3), 263-291.

[28] Parkinson, B., & Manstead, A. S. (2015). Current emotion research in social psychology: Thinking about emotions and other people. *Emotion Review*, *7*(4), 371-380.

[29] Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest*, *20*(1), 1-68.

[30] Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In *Theories of emotion* (pp. 3-33). Academic press.

[31] Kowallik, A. E., & Schweinberger, S. R. (2019). Sensor-based technology for social information processing in autism: A review. *Sensors*, *19*(21), 4787.

[32] Baig, M. Z., & Kavakli, M. (2019). A survey on psycho-physiological analysis & measurement methods in multimodal systems. *Multimodal*

Technologies and Interaction, *3*(2), 37.

[33] Diego-Mas, J. A., Fuentes-Hurtado, F., Naranjo, V., & Alcañiz, M. (2020). The influence of each facial feature on how we perceive and interpret human faces. *i-Perception*, *11*(5), 2041669520961123.

[34] Hashemi, J., Dawson, G., Carpenter, K. L., Campbell, K., Qiu, Q., Espinosa, S., ... & Sapiro, G. (2018). Computer vision analysis for quantification of autism risk behaviors. *IEEE Transactions on Affective Computing*, *12*(1), 215-226.

[35] Gregersen, T. S. (2005). Nonverbal cues: Clues to the detection of foreign language anxiety. *Foreign language annals*, *38*(3), 388-400.

[36] Hadders-Algra, M. (2022). Human face and gaze perception is highly context specific and involves bottom-up and top-down neural processing. *Neuroscience & Biobehavioral Reviews*, *132*, 304-323.

[37] Clough, S., & Duff, M. C. (2020). The role of gesture in communication and cognition: Implications for understanding and treating neurogenic communication disorders. *Frontiers in Human Neuroscience*, *14*, 323.

[38] Tripathi, A., & Thakurdesai, N. Implementation and Comparison of Facial Expression Detection and Classification Techniques. *International Journal of Computer Applications*, *975*, 8887.

[39] Gupta, A., Thakkar, K., Gandhi, V., & Narayanan, P. J. (2019). Nose, eyes and ears: Head pose estimation by locating facial keypoints. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1977-1981). IEEE.

[40] Wu, S., Du, Z., Li, W., Huang, D., & Wang, Y. (2019). Continuous emotion recognition in videos by fusing facial expression, head pose and eye gaze. In *2019 International Conference on Multimodal Interaction* (pp. 40-48).

[41] Baltrušaitis, T., Robinson, P., & Morency, L. P. (2016). Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)* (pp. 1-10). IEEE.

[42] Song, Y., & Hakoda, Y. (2018). Selective impairment of basic emotion recognition in people with autism: Discrimination thresholds for recognition of facial expressions of varying intensities. *Journal of autism and developmental disorders*, *48*, 1886-1894.

[43] Faso, D. J., Sasson, N. J., & Pinkham, A. E. (2015). Evaluating posed and evoked facial expressions of emotion from adults with autism spectrum disorder. *Journal of autism and developmental disorders*, *45*, 75-89.

[44] Meng, H., Romera-Paredes, B., & Bianchi-Berthouze, N. (2011). Emotion recognition by two view SVM_2K classifier on dynamic facial expression features. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)* (pp. 854-859). IEEE.

[45] Huang, K. Y., Wu, C. H., Su, M. H., & Kuo, Y. T. (2018). Detecting unipolar and bipolar depressive disorders from elicited speech responses using latent affective structure model. *IEEE transactions on affective computing*, *11*(3), 393-404.

[46] Ali, A., Negin, F. F., Bremond, F. F., & Thümmler, S. (2022). Video-based behavior understanding of children for objective diagnosis of autism. In *VISAPP 2022-17th International Conference on Computer Vision Theory and Applications*.

[47] Negin, F., Ozyer, B., Agahian, S., Kacdioglu, S., & Ozyer, G. T. (2021). Vision-assisted recognition of stereotype behaviors for early diagnosis of Autism Spectrum Disorders. *Neurocomputing*, *446*, 145-155.

[48] Dawood, A., Turner, S., & Perepa, P. (2019). Natural-spontaneous affective-cognitive dataset for adult students with and without asperger syndrome. *IEEE Access*, *7*, 77990-77999.

[49] Bengio, Y., Lecun, Y., & Hinton, G. (2021). Deep learning for AI. *Communications of the ACM*, *64*(7), 58-65.

[50] Choudhuri, A. R., Thakurata, B. G., Debnath, B., Ghosh, D., Maity, H., Chattopadhyay, N., & Chakraborty, R. (2022). MNIST Image Classification Using Convolutional Neural Networks. In *Modeling, Simulation and Optimization: Proceedings of CoMSO 2021* (pp. 255-266). Singapore: Springer Nature Singapore.

[51] Omran, M., & AlShemmary, E. N. (2020). An iris recognition system using deep convolutional neural network. In *Journal of Physics: Conference Series* (Vol. 1530, No. 1, p. 012159). IOP Publishing.

[52] Sharma, S., Sharma, S., & Athaiya, A. (2017). Activation functions in neural networks. *Towards Data Sci*, *6*(12), 310-316.

[53] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

[54] Vedaldi, A., & Lenc, K. (2015). Matconvnet: Convolutional neural networks for matlab. In *Proceedings of the 23rd ACM international conference on Multimedia* (pp. 689-692).

[55] Katiyar, A., Behal, S., & Singh, J. (2021). Automated defect detection in physical components using machine learning. In *2021 8th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 527-532). IEEE.

[56] Patil, A., & Rane, M. (2021). Convolutional neural networks: an overview and its applications in pattern recognition. *Information and Communication*

Technology for Intelligent Systems: Proceedings of ICTIS 2020, Volume 1, 21-30.

[57] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).

[58] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 31, No. 1).

[59] Ponti, M. A., Ribeiro, L. S. F., Nazare, T. S., Bui, T., & Collomosse, J. (2017). Everything you wanted to know about deep learning for computer vision but were afraid to ask. In *2017 30th SIBGRAPI conference on graphics, patterns and images tutorials (SIBGRAPI-T)* (pp. 17-41). IEEE.

[60] Brownlee, J. (2018). What is the Difference Between a Batch and an Epoch in a Neural Network. *Machine Learning Mastery*, *64*, 237-249.

[61] Peng, C., Xiao, T., Li, Z., Jiang, Y., Zhang, X., Jia, K., ... & Sun, J. (2018). Megdet: A large mini-batch object detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 6181-6189).

[62] Lewkowycz, A., Bahri, Y., Dyer, E., Sohl-Dickstein, J., & Gur-Ari, G. (2020). The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*.

[63] Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448-456). pmlr.

[64] Bjorck, N., Gomes, C. P., Selman, B., & Weinberger, K. Q. (2018). Understanding batch normalization. *Advances in neural information processing systems*, *31*.

[65] Xu, X., Jiang, X., Ma, C., Du, P., Li, X., Lv, S., ... & Li, L. (2020). A deep learning system to screen novel coronavirus disease 2019 pneumonia. *Engineering*, *6*(10), 1122-1129.

[66] Cook, J., & Ramadas, V. (2020). When to consult precision-recall curves. *The Stata Journal*, *20*(1), 131-148.

[67] Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*.

[68] Yao, Y., Luo, Z., Li, S., Fang, T., & Quan, L. (2018). Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 767-783).

[69] Zhang, Z. (2018). Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)* (pp. 1-2). Ieee.

[70] Ghayoumi, M., & Bansal, A. K. (2016). Unifying geometric features and facial action units for improved performance of facial expression analysis. *arXiv preprint arXiv:1606.00822*.

الخلاصة

يمكن للعاطفة التي تظهر على وجوهنا أن تحدد مشاعرنا وحالتنا العقلية ويمكن أن تؤثر بشكل مباشر على قراراتنا. يتعرض البشر للخضوع لتغيير عاطفي فيما يتعلق ببيئتهم المعيشية أو في الظروف الحالية. يمكن أن تكون هذه المشاعر اشمئزازًا أو خوفًا أو حزنًا أو فرحًا أو ترقبًا. نظرًا للتعقيد والفروق الدقيقة في تعبيرات الوجه وعلاقتها بالعواطف، يظل التحديد الدقيق لتعابير الوجه مهمة صعبة. يتميز المصابون بالتوحد بتعابير وجه غير نمطية، لذلك يجدون صعوبة في التواصل والتفاعل مع الآخرين. ونظراً للزيادة الكبيرة في أعدادهم في السنوات الأخيرة، اهتم الباحثون والعلماء بتطوير أدوات تساعدهم في التعبير عن حالتهم العاطفية. لكن لسوء الحظ، ركزت معظم الدراسات على تحليل سلوكيات التوحد باستخدام أدوات غازية. تؤدي هذه الأدوات إلى استجابة غير مرغوب فيها بسبب حساسيتها وحاجتها للتحكم في البيئة في ظل ظروف محددة. كما أن هناك نقصًا في الأعمال التي تعتمد على البيانات الطبيعية والعفوية للكشف عن سلوكيات الأشخاص المصابين بالتوحد.

لذلك، مع التطور في تقنيات التعلم العميق، طورت هذه الأطروحة نموذجًا آليًا لاكتشاف مشاعر الأشخاص المصابين بالتوحد في الوقت الفعلي باستخدام مقاطع الفيديو الديناميكية. وقد اعتمد النموذج على أهمية دور الوجه والرأس ونظرة العين في التفاعل مع الآخرين لتكون مدخلاً لهذا النموذج .المساهمة الجديدة لهذا العمل هي التحقيق في اختلافات تعبيرات الوجه بين الأشخاص المصابين بالتوحد وأولئك الذين لا يعانون من التوحد، أثناء التعبير عن نفس المشاعر.

ساهم استخدام خوارزمية Googlenet في تحقيق نتائج مهمة في استخلاص الميزات والتنبؤ بالعواطف. حيث بلغت دقة تدريب النموذج والتحقق من صحته ٩٩.٨٨% و٩٨.٥٤% على التوالي، كما أثبت النموذج كفاءته تجاه البيانات غير المرئية بدقة بلغت ٩٧.٥٤%.

جمهورية العراق

وزارة التعليم العالي و البحث العلمي

جامعة بابل – كلية العلوم للبنات

قسم علوم الحاسوب

تمييز العواطف من تعابير الوجه للأشخاص المصابين بالتوحد باستخدام التعلم العميق

رسالة مقدمة الى

مجلس كلية العلوم للبنات-جامعة بابل

وهي جزء من متطلبات نيل درجة الماجستير في علوم الحاسبات

من قبل

ميس علي شاكر

بإشراف

أ.م.د. امنة عطية داوّد

٢٠٢٣ ١٤٤٥