"MULTIMODAL EMOTION RECOGNITION IN EDUCATIONAL SETTINGS USING AI AND FORENSIC PSYCHOLOGY"

A Thesis

Submitted towards the Requirements for the Award of Degree of

DOCTOR OF PHILOSOPHY

IN

COMPUTER SCIENCE AND ENGINEERING

Under the Faculty of Engineering and Technology

Submitted By

Mr. AMIT GHOSH

(Enrollment No. 220101199)

Under the Supervision of

Dr. SANTANU SIKDAR

Faculty of Engineering and Technology

P.K. UNIVERSITY



2026

P.K. University

NH-27, Vill, Thanra (P.O. Dinara)

Shivpuri, M.P. India-473665



This Piece of Work
is

Dedicated
To

My Beloved Parents

Whose blessings & love, brought me to this level.....

Amit Ghosh .....

P.K.UNIVERSITY
SHIVPURI (M.P.)

University Established Under section 2F of UGC ACT 1956 Vide M.P. Government Act No17 of 2015

CERTIFICATE OF THE SUPERVISOR

This is to certify that the work entitled **"MULTIMODAL EMOTION RECOGNITION IN EDUCATIONAL SETTINGS USING AI AND FORENSIC PSYCHOLOGY"** is a piece of research Work done by **Mr. AMIT GHOSH** Under My Guidance and Supervision for the degree of Doctor of Philosophy of University (M.P) India.

I certify That the candidate has put an attendance of more than 240 day with me.

To the best of my knowledge and belief the thesis:

I –Embodies the work of the candidate himself / herself.

II –Has duly been completed.

III – Fulfil the requirement of the ordinance relating to the Ph.D. degree of the University.

Date:…… **Signature of the Supervisor**

DECLARATION BY THE CANDIDATE

I declare that the thesis entitled **"MULTIMODAL EMOTION RECOGNITION IN EDUCATIONAL SETTINGS USING AI AND FORENSIC PSYCHOLOGY"** is my own work conducted under the supervision of **Dr. SANTANU SIKDAR, Department of Computer Science and Engineering, Faculty of Engineering and Technology, at P.K. University, Shivpuri, (M.P.)**

Approved by Research Degree Committee. I have put more than 240 days of attendance with supervisor at the center.

I further declare that to the best of my knowledge the thesis does not contain my part of any work has been submitted for the award of any degree either in this University or in any other University Without proper citation.

Date:-…… **Signature of the candidate**

Place: P.K. University, Shivpuri, (M.P.)

P.K.UNIVERSITY
SHIVPURI (M.P.)

University Established Under section 2F of UGC ACT 1956 Vide M.P. Government Act No17 of 2015

COPYRIGHT TRANSFER CERTIFICATE

Title of the Thesis: **"MULTIMODAL EMOTION RECOGNITION IN EDUCATIONAL SETTINGS USING AI AND FORENSIC PSYCHOLOGY"** Candidate's Name: **Mr. AMIT GHOSH**

COPY RIGHT TRANSFER

The undersigned hereby assigns to the P.K. University, Shivpuri (M.P.) all copyrights that exist inand for the above thesis submitted for the award of the Ph.D. Degree.

Date:…….. **Name of the scholar**

Note: However, the author may reproduce/publish or authorize others to reproduce, material extracted verbatim from the thesis or derivative of the thesis for author's personal use provided that the source and the University's copyright notice are indicated.

P.K.UNIVERSITY
SHIVPURI (M.P.)

University Established Under section 2F of UGC ACT 1956 Vide M.P. Government Act No17 of 2015

FORWARDING LETTER OF HEAD OF INSTITUTION

The Ph.D. thesis entitled **"MULTIMODAL EMOTION RECOGNITION IN EDUCATIONAL SETTINGS USING AI AND FORENSIC PSYCHOLOGY"** Submitted by **Mr. AMIT GHOSH Enrollment No.:- 220101199** University in six copies. The candidate has paid the necessary fees and there are No-dues outstanding against her.

Name:- Dr. Ashish Vishwakarma Seal……

Date:……

Place:- P.K. University, Shivpuri, (M.P.)

(Signature of Head of Institution where the Candidate was registered for Ph.D. degree)

Signature of the Supervisor

Date: ……..

Place:- P.K. University, Shivpuri, (M.P.)

DrillBit

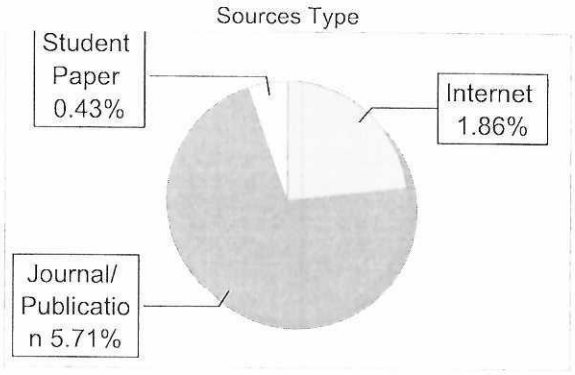
The Report is Generated by DrillBit Plagiarism Detection Software

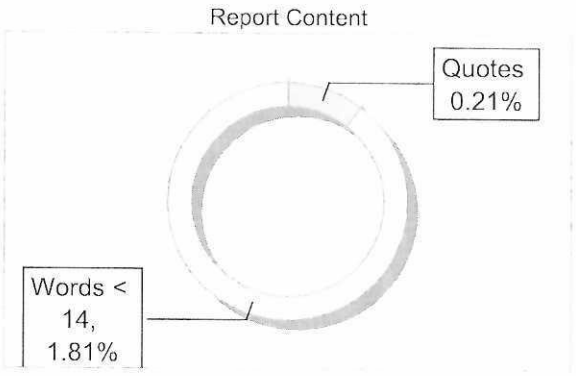
Submission Information

Author Name	Amit Ghosh
Title	MULTIMODAL EMOTION RECOGNITION IN EDUCATIONAL SETTINGS USING AI AND FORENSIC PSYCHOLOGY
Paper/Submission ID	5564980
Submitted by	library.pku@gmail.com
Submission Date	2026-04-29 11:36:13
Total Pages, Total Words	194, 48759
Document type	Thesis Chapter

Result Information

Similarity **8 %**





Exclude Information

Quotes	Not Excluded
References/Bibliography	Not Excluded
Source: Excluded < 14 Words	Not Excluded
Excluded Source	0 %
Excluded Phrases	Not Excluded

Database Selection

Language	English
Student Papers	Yes
Journals & publishers	Yes
Internet or Web	Yes
Institution Repository	Yes

LIBRARIAN
P.K. University
Shivpuri (M.P.)

P.K.University
Shivpuri (M.P.)

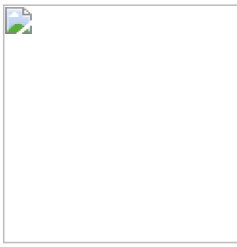
Result - September-2023

Course:	Ph.D. in Computer Science & Engineering	Semester:	First Semester
Enrollment No:	220101199	Institute Name:	Faculty of Engineering & Technology
Student Name:	AMIT GHOSH	Roll No:	226144204
Father's Name:	DILIP GHOSH	Session:	2022-23

S.No.	Subject Code	Subject Name	Credit	Grade	Grade Point	Credit Point
1	PHD101	RESEARCH METHODOLOGY	4	B+	8	32
2	PHD102	COMPUTER SCIENCE & ENGINEERING	6	B+	8	48
3	PHD103	RESEARCH AND PUBLICATION ETHICS	2	B+	8	16
		Total	12			96

SGPA:	96 / 12	8
CGPA:		8
Result:		Pass

* Grade in Repeat Examination



ACKNOWLEDGEMENT

The Acknowledgement of a research endeavour is an opportunity to express appreciation for the joint effort of those who have been sources of motivation and support, and who have contributed invaluable knowledge towards achieving success in the pursuit of knowledge. With utmost humility and pride, I reflect on the inspiring individuals who have propelled my work forward, and I am filled with immense pleasure, gratitude and thanks for their unwavering dedication.

First, for most I would like to thank almighty god for perseverance, courage and strength bestowed on me to undertake this project.

For most, I extend my deepest appreciation to my esteemed Supervisor, Dr. **Santanu Sikdar,** Department of Computer Science & Engineering, Faculty of Engineering & Technology , P.K. University, Shivpuri, (M.P.) without whom this work would not have come to fruition. His exceptional guidance, unwavering support kind cooperation and provision of a conductive work environmental have been invaluable beyond measure. I remain indebted to her instilling in me a relentless pursuit of excellence, a strong sense of honesty, and respect for ethical principles that govern our profession.

Furthermore, I extend my sincere thanks to the **Mr. Jagdeesh Prasad Sharma** Chancellor, P.K University Shivpuri (M.P) for providing the necessary infrastructure and facilities that were essential during my research work. My profound sense of gratitude, faith, and we also goes to **Prof. (Dr.) Yogesh Chandra Dubey** Vice Chancellor, P.K University Shivpuri (M.P) for his generous and valuable support throughout my research.I express my gratitude to **Prof. (Dr.) Mukesh Chansoriya** ProVice -Chancellor, P.K University Shivpuri (M.P) for his generous and valuable support throughout my research.

I take this pleasant opportunity to express my deep sense of gratitude and reverences to his holiness **Dr. Jitendra Kumar Mishra** Director; P.K

University Shivpuri (M.P) for his constant encouragement and providing mewith therequired facilities that enabledmeto complete my project work.

Dr. Jitendra Malik Registrar, P.K University, Shivpuri (M.P.), who has inspired me with his encouraging guidance and support right from the beginning of this work. in spite of his busy schedule. I am short of words to thank him for his affectionate behavior and patience throughout the duration of research work.

I express my gratitude to **Dr. Aiman Fatima,** Dean Academics, of P.K. University, Shivpuri, M.P.

I would also like to extend my gratitude to and deep appreciation to respected **Dr. Ashish Vishwakarma ,** Dean Research of P.K. University, M.P.

I owe a great deal of application and gratitude to entire faculty members of Faculty of Engineering Technology, P.K. University, especially **Dr. Monika Tripathy** and all non-teaching staff of department of Computer Science & Engineering without their support it would have been difficult to shape my research work.

I also express my gratitude to **Miss. Nisha Yadav,** Librarian of P.K University along with **Mr. Anand Kumar,** Assistant Librarian and other staff of P.K University. I am also thankful to **Mr. Pankaj Singh, Mr. Alok Khare,** and **Mr. Prushottam Yadav** I.T. Cell of P.K. University.

I have no words to express my gratitude to my beloved family for their love, support and constant encouragement during the entire course of my research work.

Last but not least, I would like to honestly appreciate the fact that it is my pleasure to have an opportunity to pursue my higher education at P.K. University, Shivpuri (M.P). I owe my permanent debt towards it, this is having the special place in my heart for shaping my care.

I must also express my love and affection to my family members for his understanding and unwavering love during my research tenure, Lastly, I am highly grateful to God, the almighty, for giving me courage and strength to face the most difficult times of life and for answering my prayers in a special way. I remain grateful for all that I have achieved in my life thus far.

Date:……

Place:…….. **(Amit Ghosh)**

Table of Contents

CHAPTER – 1 1
INTRODUCTION 1
1.1 BACKGROUND OF THE STUDY 1
1.2 EMOTION 2
1.2.1 Emotion, Feeling, Moods, Affect, Sentiments, And Emotional Dimension 3
1.3 HUMAN EMOTION 12
1.4 EVOLUTION OF HUMAN COMPUTER INTERACTIONS 16
1.5 ROLE OF EMOTIONS IN HCI 17
1.6 AFFECTIVE COMPUTING 18
1.6.1 Relevant Case Studies and Applications of Affective Computing 19
1.6.2 Potential Concerns 21
1.7 DEVELOPING GUIDELINES FOR EXAMINING AUTOMATED EMOTION RECOGNITION'S STATUS AT THIS MOMENT 22
1.8 MULTIMODAL EMOTION RECOGNITION 24
1.8.1 The role of multiple modalities 25
1.8.2 Current approach 28
1.8.3 Research Issues and Challenges 34
1.9 FORENSIC PSYCHOLOGY 39
1.9.1 Applications of Forensic Psychology 40
1.9.2 Challenges of Investigations in Forensic Psychology 43
1.9.3 Forensic Psychology Investigation Methods 44
1.11 MOTIVATIONS 46
1.12 CONTRIBUTIONS OF THESIS 47
1.13 OBJECTIVE 48
1.14 SCOPE OF THE STUDY 48
1.15 LIMITATIONS OF THE STUDY 49
1.16 RESEARCH GAP 49
CHAPTER – 2 50
LITERATURE REVIEW 50
2.1 INTRODUCTION 50
2.2 FACIAL EMOTION RECOGNITION TECHNIQUES 53
2.3 SPEECH / AUDIO EMOTION RECOGNITION TECHNIQUES 56
2.4 MULTIMODAL EMOTION RECOGNITION APPROACHES 60
2.5 OPTIMIZATION TECHNIQUES IN EMOTION RECOGNITION 64

2.6 SUMMARY OF RESEARCH GAPS **70**
CHAPTER – 3 74
DESIGN AND METHODS 74
3.1 OVERVIEW **74**
3.2 PROPOSED SYSTEM ARCHITECTURE **75**
3.3 EXPERIMENTAL SETUP **79**
3.4 DETAILS OF DEAP DATASET **80**
3.5 DATA COLLECTION **82**
3.6 IMAGE PREPROCESSING **82**
3.7 FACE DETECTION **83**
3.8 PRE TRAINED MODELS **83**
3.9 IMPLEMENTATION TOOLS **85**
3.10 DEVELOPMENT **86**
3.11 EVALUATION METHODS **86**
3.11.1 Confusion Matrix 86
3.12 MULTIMODAL EMOTION RECOGNITION FOR HUMAN- COMPUTER INTERACTION **89**
3.13 INFORMATION FUSION **94**
3.14 FACIAL EMOTION RECOGNITION USING MULTIMODAL SIGNALS **95**
3.14.1 Gathering of input video 96
3.14.2 DBN-based facial emotion recognition 101
3.14.3 Deep RNN-based Facial Emotion Recognition 106
3.15 ALGORITHM FOR MULTIMODAL EMOTION RECOGNITION USING DEEP LEARNING NETWORKS **115**
3.16 FLOW CHART OF PROPOSED BROA-DEEP NN **117**
3.17 SYSTEM IMPLEMENTATION **119**
CHAPTER – 4 125
EXPERIMENTAL RESULTS AND DISCUSSIONS 125
4.1 OVERVIEW **125**
4.2 DATASET OUTLOOK **125**
4.3 EXPERIMENTAL SETTING **126**
4.3.1 Preprocessing the dataset 127
4.3.2 Data Preparation 139
4.3.3 Load the Validation and Training Datasets 140
4.3.4 Base Model assessment 141
4.3.5 Assessment of Defining the Architecture of the Model 142
4.3.6 Functions for Callbacks 143

4.3.7 Models for trains 144
4.4 RESULTS OF THE EXPERIMENT 145
4.4.1 Confusion Matrix Analysis 145
4.4.2 Loss and accuracy 149
4.4.3 Precision values for every emotion 151
4.5 F1-SCORE 153
4.5.1 Reports on the classification of every class or emotion 153
4.5.2 ROC Curve Analysis 156
4.6 STRATEGY ANALYSIS 157
4.7 ANALYSIS OF TOOLS USED FOR IMPLEMENTATION 159
4.8 DATA ANALYSIS FOR TRAINING 162
4.8.1 Analysis with Different Hidden Neurons 162
4.8.2 Analysis with Different Rates of Learning 167
4.8.3 A feature Analysis 172
4.9 PERFORMANCE ANALYSIS 177
4.9.1 Accuracy 177
4.9.2 False Acceptance Rate (FAR) 177
4.9.3 False Rejection Rate (FRR) 178
4.10 COMPARISON WITH EXISTING METHODS 178
4.10.1 FAR of Proposed BROA-Deep NN with the Existing Methods 179
4.10.2 FRR of the Proposed BROA-Deep NN with the Existing Methods 181
4.11 OVERVIEW OF DATA SOURCES 182
4.11.1 Facial Emotion Dataset-FER-2013 183
4.11.2 Audio Emotion Dataset – RAVDESS 183
4.11.3 Multimodal / Physiological Dataset – DEAP 184
4.11.4 Data Usage and Ethical Compliance 185
CHAPTER - 5 186
CONCLUSION AND FUTURE SCOPES 186
5.1 CONCLUSION 186
5.2 LIMITATIONS OF THE STUDY 190
5.3 MAJOR CONTRIBUTIONS OF THE STUDY 190
5.4 APPLICATION DOMAINS OF THE PROPOSED MODEL 191
5.5 FUTURE SCOPE 191
REFERENCES 194

ABSTRACT

Understanding and detecting human emotions by combining data from several sources, such as speech, text, facial expressions, and physiological signs, is known as multimodal emotion identification. Ubiquitous Computing describes how computers and their apps have become an integral part of modern living. Users' interactions with computers are increasingly resembling those between humans.

The goal of incorporating emotion detection into HCI is to make the interaction more natural and effortless. This study provides a concise overview of where multimodal emotion detection is at within the Human-Computer Interaction (HCI) discipline. We talk about recognizing emotions. as an emerging field of study, its potential uses in HCI, and the difficulties and rewards of this field. Finally, a proposal for the creation of a multimodal emotion detection analyzer is included, following the presentation of the research on information fusion and human-computer interaction. The primary objective of this work was to select a suitable model for face emotion identification using transfer learning techniques. The dependability of the models used is the main focus of the assessment.

In particular, in order to ascertain if interpolation yields superior outcomes, the researchers wish to conduct an experimental assessment of the interpolation technique's effectiveness in upscaling lower-resolution pictures.

To demonstrate that interpolation increases the precision of models used for image processing tasks, especially the identification of facial emotions, this research will examine its effects on model performance and picture quality in a systematic way.

This By methodically examining its effects on picture quality and model performance, the study seeks to empirically demonstrate how well interpolation works to increase the accuracy of models used in image processing tasks, particularly for face emotion recognition.

CHAPTER – 1
INTRODUCTION

1.1 BACKGROUND OF THE STUDY

It was in the prehistoric period that man first fashioned the wheel by shaping stone. As far as machine development goes, this was the first leap. After that, people started to see the value in machines, and today there are a lot of them all over man. There are many devices with which we interact, such as watches, TVs, mobile phones, computers, cars, etc. These devices would have been unthinkable a few centuries ago, yet now they are indispensable. Machines impact our lives in many ways; some examples are aircraft, rockets, satellites, high-tech medical equipment, and nuclear technology.

The amount to which we rely on machines has been growing lately. Machines are increasingly being outfitted with more intuitive user interfaces to facilitate faster and more efficient interaction. We are now totally reliant on machines because of how quickly technology has integrated human activities with them. An excellent illustration of this kind of reliance is the ubiquitous computer, which has spread from its original domain in labs to almost every facet of contemporary existence.

Technology's objective is to facilitate our daily life. Smaller personal devices like PDAs, smartphones, and palmtops have started to augment and even replace larger computing platforms like personal computers in today's technological age. Future computers will be smaller, cheaper, and have more add-on functions, all because of this trend.

Emotional intelligence and management are seen as fundamental to success in the modern workplace, business, and consumer goods (Norman, 2004) [1]. In terms of usability, users care about more than just features; they want experiences that cater to their expectations, emotions, and the purposes of their interactions. So, how end users perceive a system's usability is the deciding element, which presents ongoing difficulties for HCI researchers.

1.2 EMOTION

People frequently say that our emotions are our "motus anima," or what really motivates us. The wide range of emotions and the inspiration they provide continue to have a significant influence, no matter how much we try to be logical and reasonable.

Emotions are multi-faceted mental and physical states that may be set off by several internal and external elements. They consist of a broad range of subjective experiences that might impact one's ideas, actions, and physiological reactions, including emotions, moods, and affective states. As a result, the scientific community has put out an increasing number of definitions for the concept of emotion, despite the fact that it is a widely used and seemingly simple phrase.

The sheer number of these definitions has increased to an absurd level, rendering any effort at counting them futile. One example is the 1981 study by Kleinginna and Kleinginna, which examined more than 100 definitions. Therefore, while it's hard to come up with a complete description, we may learn about the many viewpoints on this subject by looking at some of the most important definitions proposed by psychologists and others. So, by looking at the most influential definitions and models, we may see how people like psychologists have tackled this issue.

Emotion, according to the theories, is a multi-faceted concept. Everything from our thoughts and actions to our decisions and interactions with others is greatly impacted by our emotions. Insights like this may motivate us to take action, help us communicate with others, and reveal important details about our environment. Our ability to remember things, focus our attention, and make judgments—even ones fueled by emotion—are all impacted by our emotional states.

When people are feeling emotionally stable, they are better able to control their reactions, make sound judgments, and manage social situations. Emotional disturbances and disorders may have a devastating effect on a person's ability to function and overall happiness, thus they are also vital to mental health and wellness. In an effort to decipher the nature, function, and history of emotion, a plethora of competing hypotheses have evolved. Many modern researchers continue to rely on classic ideas as initial frameworks,

despite the fact that they have been the target of criticism.

The James-Lange Theory, which proposed that our physiological reactions to outside stimuli are the root cause of our emotions, was the first in this line of thinking. Later, the Cannon-Bard Theory put this idea to the test by proposing that, as the thalamus analyzes inputs, the brain experiences both emotional and physiological reactions concurrently but separately.

By integrating physiological arousal with cognitive interpretation, the Schachter-Singer model improved upon the James-Lange hypothesis. Individuals have physiological arousal in reaction to stimuli, which is then interpreted by contextual signals and labelled emotionally, according to this paradigm.

Affective Events Theory, which delves into the temporal dimension of emotional responses, and contemporary Cognitive Appraisal Theories like Cognitive-Mediational Theory put forward by Richard Lazarus are examples of more recent contributions. The impact of cognitive assessment on the formation of emotional experiences is emphasized by these theories.

A number of terms related to emotions have their origins in older philosophical frameworks. These include sentiment, emotion, feeling, mood, affect, and emotional dimensions. Although these phrases are synonymous, they have different meanings.

1.2.1 Emotion, Feeling, Moods, Affect, Sentiments, And Emotional Dimension

Originating with ancient thinkers like Socrates, the use of these terms first appeared in scientific corpora in the 18th century. There is a tight relationship between these phrases, yet they signify different things and have different conceptions.

A complicated and often involuntary psychological and physiological reaction to a specific stimulus is what we call an emotion. It is generally believed to be both universal and physiologically based, and it causes changes in the body including the heart rate, facial expressions, and hormone release.

The term "feeling" refers to the subjective or conscious knowledge of an emotional state. Perceiving and understanding the internal and external changes that take place as a result of an emotional reaction is an important part of this process. Feelings are the subjective

experiences and interpretations of an emotion, as opposed to emotions, which are more general processes that include physiological reactions.

A mood is a condition of prolonged emotional distress that does not have a beginning or conclusion. A person's emotional state and their perception and reaction to events might be influenced by these less powerful feelings, which can last for hours, days, or even longer. Affect is a powerful, unstructured, and unconscious feeling that exists either before or beyond the threshold of consciousness. It trains the body to react strongly to stimuli by taking into account settings larger than the individual stimuli themselves.

Emotions, feelings, and beliefs create sentiments, which are complex and long-lasting emotional attitudes or views regarding people, things, or ideas. One aspect that highlights their social origins and effect is the presence of long-lasting social ties that elicit emotions.

Dimensions of emotion include valence and arousal. The term "arousal" describes the degree of mental and physical activity linked with an emotional state, which may range from relaxed to agitated. The positive or negative aspect of an emotion is represented by its valence, which may go from pleasant to unpleasant.

B. Emotion Models

Emotions are triggered by different parts of the brain. Reactive, hormonal, and automatic responses are the three main types of emotional reactions. From a psychological perspective, Emotions are reactions to outside stimuli linked to particular shifts in one's physiological make-up. The two main ways that researchers look into what emotions are all about are the discrete technique and the multidimensional approach.

Discrete emotions theory classifies emotions into different groups, with unique traits in terms of behavior, cognition, and psychology. Those who support this view argue that there are a handful of basic emotions that are understood by people of all cultures, and these feelings may be pleasant or bad. Feelings of joy, grief, surprise, dread, and disgust are among these fundamental emotions. As an extension of this idea, Robert Plutchik developed a thorough Plutchik's wheel of emotions is a model of human emotions. Eight primary emotions are represented in this model: fear, joy, sadness, trust, anger, surprise, anticipation, and disgust. Position on the wheel determines the strength of secondary

emotions, which combine these main ones; intensity increases in the center and decreases towards the perimeter. A synopsis of Plutchik's emotional wheel is shown in Figure. 1.1.

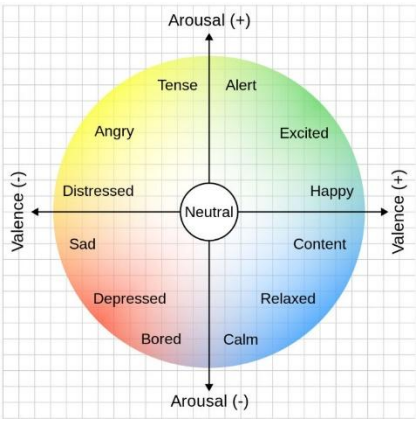
Personal history, cultural background, and individual variances are only a few of the many aspects that impact an individual's emotional experiences, according to the Multidimensional Emotions Theory. A more thorough examination of emotional states is made possible by this framework, which offers a more profound comprehension of emotions. Within this framework, there is a categorization of emotional space models into two- and three-dimensional varieties. Valence (good or bad) and arousal (high or low activity) are the two dimensions upon which the 2D model classifies emotions. Known variously as the Circumplex Model of emotion and Russell's 2D emotional space model, this mapping is shown.

Along with valence and arousal, the 3D model also includes dominance, which is defined as the sensation of being in charge or being dominated. This multi-dimensional mapping is shown by the three-dimensional Mehrabian and Russell created the emotional space paradigm. [2]



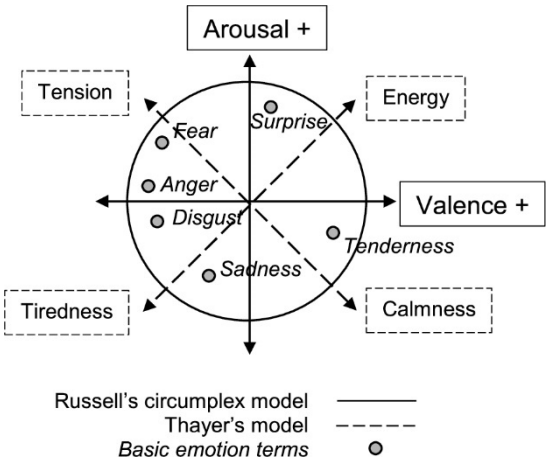
Source: https://www.researchgate.net/figure/Plutchiks-Wheel-of-Emotions_fig1_349016413

Figure 1.1: Plutchik's emotional wheel



Source: https://commons.wikimedia.org/wiki/File:Circumplex_model_of_emotion.svg

Figure 1.2: Circumplex 2D VA emotion model Emotional model



Source: https://informatica.vu.lt/journal/INFORMATICA/article/1182/text

Figure 1.3: 3D representation of VAD emotion.

C. Expression of Emotions

Physiological reactions, facial expressions, body language, voice intonation, and gestures are just a few ways that humans convey their emotions. There are two ways in which emotions may be shown: inwardly and outwardly. The two-stage model of emotional expression proposed by Darwin posits that emotions serve to both respond to our environment and communicate social signals. Noting commonalities in production and interpretation across cultures, Keltner et al. emphasize the social nature of emotions. Emotions exist and may be expressed in ways we aren't always aware of, and our brains can react to previous experiences even when we're not consciously aware of them; this is

something Winkielman and Berridge investigate.

Emotional regulation theory explains how to keep one's emotions under check in various situations. Emotional intelligence encompasses a wide range of techniques for comprehension, monitoring, and modification of emotional states, including the ability to identify and interpret emotions, identify and mitigate their triggers, and enhance good emotions while dampening negative ones. The ability to control one's emotions is frequently seen as the natural progression from expressing them to expressing them creatively and appreciating beauty.

The fact that different people experience different emotions via different senses was brought to light in Charles Darwin's early writings, which described more than 40 emotional states. Paul Ekman's studies highlighted the universality of facial expressions for fundamental Feelings include joy, grief, rage, fear, surprise, and contempt, highlighting both their biological foundation and their ability to be recognized across cultures. Modern affective science and emotional regulation studies owe a debt to this pioneering work in comprehending the nuanced nature of human emotion expression. The significance of body language and vocal intonation in expressing emotions is emphasized by Albert Mehrabian's studies on nonverbal communication, specifically his "7-38-55 rule," which is often misunderstood and questioned.

The emotional expression finds expression in a variety of forms, going beyond the realm of physiological motions. In his writings, Langer explores the ways in which visual art, literature, and music convey feelings. Changes in heart rate, voice intonation, facial expressions, and sleep patterns are all signs of emotional state. Expressions of art, language, and communication strategies all function as conduits for emotions. Expressions of emotion may be explored in visual art, music, painting, and drawing. Often seen as secondary or social forms of emotional expression, these modalities may overlap with bodily manifestations, such in the theater or communication made easier by laughing. Emotional regulation is a common process that helps to improve these manifestations.

Based on the two main types of emotional expression—physical and non-verbal expression and communicative and artistic expression—Figure. 1.4 shows the two main types of emotional expression and how they're expressed.

Emotions may cause physiological reactions, which manifest as changes in the body's temperature, perspiration rate, odor, shaking, and blood pressure. Some examples of these include:

a. **Confrontational displays:** They are well-known for conveying a wide range of emotions. We show our emotions via our faces in unique ways whether we're happy, sad, angry, afraid, surprised, or disgusted.

b. **How one stands:** Posture, gestures, and movements all constitute body language, and they may express emotions. As an example, if someone feels angry, they may clench their fists, and if they are sad, they may stoop their shoulders.

c. **One's tone of voice:** How quickly and how high we speak might be a window into how we're feeling. If you're nervous or excited, your voice may be too high-pitched, and if you're unhappy or bored, it could be too sluggish and monotonous.

d. **Pattern of heart rate:** The cardiovascular system's dynamic reaction to emotional events is represented by the heart rate's rhythmic variation. throughout time. An individual's heart rate, for instance, rises in response to excitement or arousal and falls in response to rest.

e. **Shifts in thinking:** Our ideas and cognitive processes are impacted by our emotions. Feeling happy may boost your creativity and problem-solving skills, but anxiety may cause your thoughts to race and make it hard to concentrate.

f. **Responses in behavior:** Crying, laughing, aggressiveness, or avoidance are among ways that emotions manifest in behavior. There is a generalization of these actions; for example, laughing may sometimes be a sign of nervousness or wrath.

g. **Sleep schedule:** Moods and thoughts may influence how well you sleep. When you're feeling down, anxious, or stressed out, it could be hard to get a good night's rest.

h. **Eye activity:** Changes in pupil size, eye contact patterns, and fixations may all be indicators of emotional states. For instance, excitement levels are correlated with dilated pupils.

i. **How the body breathes:** A person's breathing rate and pattern might fluctuate in response to their emotions. Breathing more quickly and shallowly can be a sign of stress or excitement, whilst breathing more slowly and deeply might be a sign of calm.

Aside from the obvious ways in which our bodies show our emotions, one of the most important ways is via our relationships with others, especially at the more abstract, imaginative, and creative levels of human connection. Painting, sketching, sculpture, photography, and digital art are all forms of visual expression that fall under the umbrella of artistic expression. Color, composition, shape, texture, and symbolism are some of the tools artists employ to communicate with audiences and elicit reactions. By appealing to the universal language of pictures and aesthetics, creative expression enables people to express nuanced feelings that can be hard to put into words. This includes:

i. Literature and the written word

Authors artistically express feelings, ideas, and life events via various forms of written expression such as prose, poetry, essays, and letters. In this way, nuanced feelings may be expressed and understood.

ii. The performing arts

Through its pace, harmony, rhythm, and melody, music evokes a wide range of feelings. An array of voice styles and instruments may elicit a wide range of feelings.

iii. Art (painting, drawing)

The expressive potential of these visual arts is immense. Artists convey a wide spectrum of emotions and life experiences via the use of shape, color, and line.

iv. Visual design

Visual design, which aims to evoke feelings via the use of form, texture, color theory, and composition. Visual design has the power to evoke emotions and provide aesthetically pleasing experiences.

v. The use of words

Emotions are conveyed via language, whether it be spoken or written. Communicating one's feelings aloud may help people connect with others, empathize, and provide emotional support.

D. Emotion recognition

Emotion recognition requires perceptual and cognitive processing of human emotional states conveyed by spoken, nonverbal, and physiological signs. Either people or automated systems are capable of doing this.

i. None-automated emotion recognition

People have an innate talent for reading facial expressions, vocal intonation, body language, and context to deduce an individual's emotional state. Intuitively, people are capable of accurately determining other people's emotional states via a series of intricate cognitive and perceptual processes. Although detecting emotions in people is usually a complex and natural process, scientists have also created more systematic ways to do it. For example, there is The Self-Assessment Manikin (SAM) examines subjective emotional experiences, the Geneva Emotion Wheel categorizes various emotional states, and the Facial Action Coding System (FACS) analyzes facial expressions. Also, in order to scientifically evaluate emotional experiences and moods, a number of popular questionnaires and scales have been created. One tool that uses self-report assessments of different emotional states to evaluate affectivity is PANAS stands for Positive and Negative Affect Schedule. The State-Trait Anxiety Inventory (STAI) and the Beck Depression Inventory (BDI) are two tools used to quantify anxiety. They gauge the degree of depressive symptoms. In addition, the POMS assesses fleeting emotional states along several dimensions.

ii. Automated Emotion Recognition

The term "automated emotional regulation" (ER) refers to the method by which computers can detect and understand human emotions from a variety of signals. Many fields are

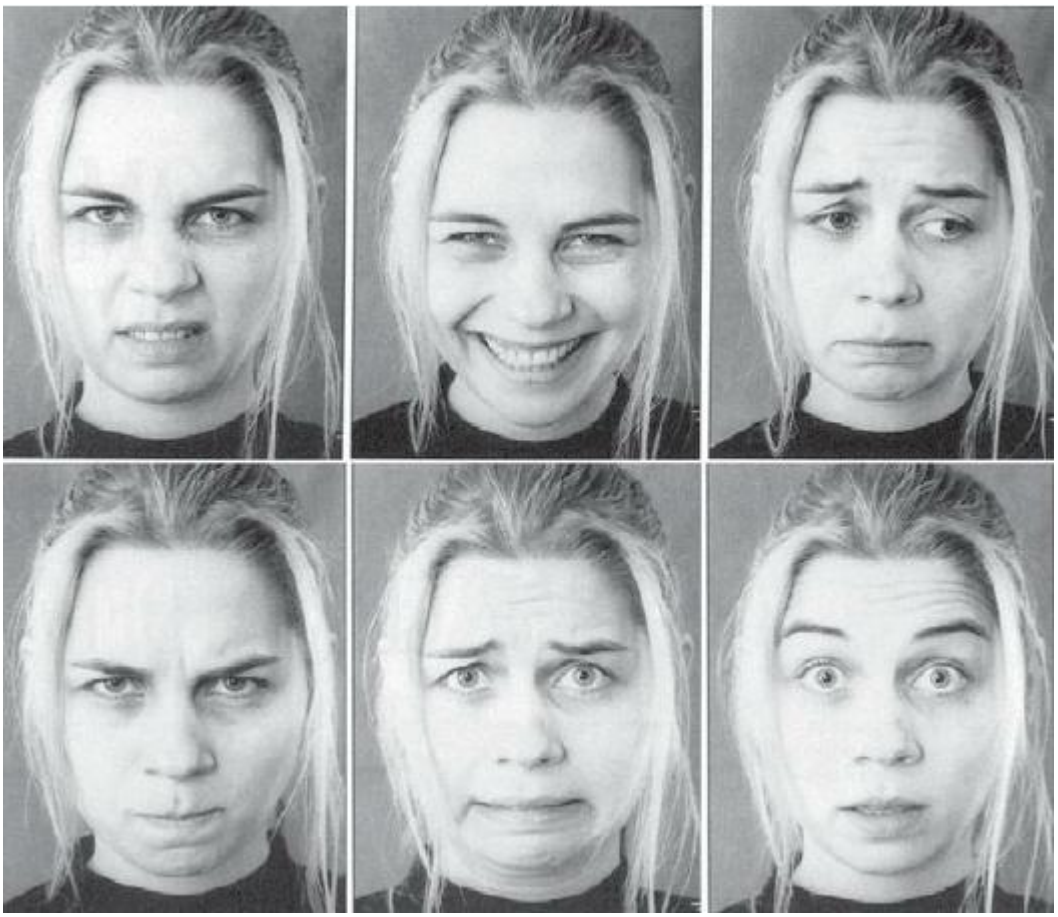
touched by it, including psychology, education, health, and entertainment, as well as human-computer interaction. This field presents difficulties and calls for a variety of approaches, such as natural language processing, computer vision, machine learning, and multimodal data analysis, given the complexity and diversity of human emotions. The ultimate objective of integrating intelligent systems into a variety of devices and applications is emotion detection.

The first attempts to study human ER were monomodal, meaning they only looked for one modality of expression or detection. Because datasets are easily accessible and computer approaches are constantly improving, the most researched modality in this context is face recognition, followed by voice and text emotion detection.

In a bimodal system, for example, visual and auditory data are usually combined (A-V). To deduce people's emotional states, a bimodal system can use both auditory and visual cues, including facial expressions and the tone of speech. When a combination of two modalities is enough for the job at hand, bimodal systems are used because they provide a reasonably straightforward way to integrate the two sources of information.

To identify human emotions, a multimodal approach takes into account not just verbal cues but also nonverbal cues such as body language, facial expressions, and physiological data. Human emotions, their expression, and their detection are complicated and context-dependent, making this a difficult endeavor. It is common practice to use machine learning, computer vision, and natural language processing (NLP) methods to combine and evaluate these many modalities in an effort to tackle this difficulty; nevertheless, these approaches may further complicate matters. Humans usually communicate their feelings across numerous modalities, thus if multimodal detection can conquer its complexity, it might provide the most accurate way to identify emotions. Whether an ER approach is unimodal, bimodal, or multimodal is dependent on a number of aspects, including the application's context, the complexity of the emotional states that need to be identified, and the resources that are available. A breakdown of each method follows:

Picking the right modality for emotion recognition has its own challenges, in addition to deciding between unimodal, bimodal, and multimodal approaches. [3]



Source: *Bennett, Casey & Sabanovic, S. (2014) https://www.researchgate.net/publication/267391519_Deriving_Minimal_Features_for_Human-Like_Facial_Expressions_in_Robotic_Faces?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6Il9kaXJlY3QiLCJwYWdlIjoiX2RpcmVjdCJ9fQ*

Figure 1.4: Human emotional expression.

1.3 HUMAN EMOTION

Ideas, the intensity of positive or negative experiences, and behavioral responses all impact an individual's emotional state. Human emotion does not have a single, widely recognized definition at this time. Over the last 20 years, researchers in domains as diverse as computer science, neuroscience, and psychology have made great strides in understanding human emotion. The literature is subjective when it comes to defining human feeling. Emotions may be better understood by using theories from both cognitive neuroscience and physiological research. According to the physiological view, feelings originate in the human body. It is possible that the alteration in physiological reaction is due to the outside incident. A common example is the widespread fear of snakes. According to cognitive

neuro-science theory, it is the brain's own processes that are responsible for producing an emotional response. The prefrontal cortex is responsible for controlling emotions and making decisions, as well as storing information about oneself. Thoughts and other mental pursuits are linked to emotions. As an example, seeing the snake makes individuals feel threatened and causes them to become scared. Emotions are defined as changes in physiological and cognitive reactions to stimuli, such as pleasure or pain.

- **Emotional component procedure**

Evoking human emotion is described by the component process, which requires a series of events whose synchronization. Psychologists agree that there are a lot of moving parts when it comes to making people feel anything. For a thorough evaluation of human emotion, Klaus Scherer [4] lays forth the five-step procedure. Mental evaluation, physical manifestations, inclinations to act, and subjective emotions are all part of it. These parts provide a multi-faceted description of human emotion. Individuals' subjective explanations of what's happening make up the cognitive appraisal component.

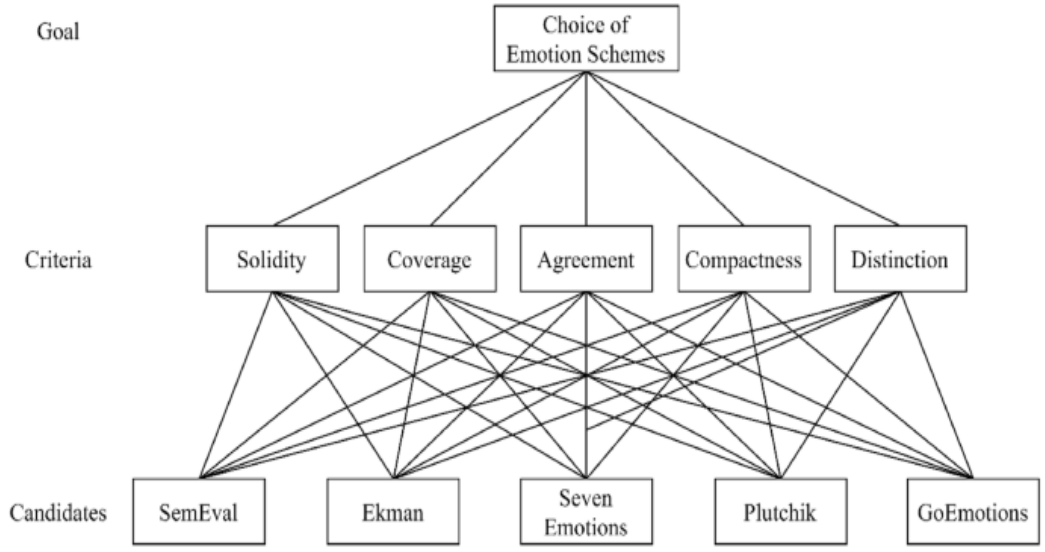
People often associate snakes with danger and fear when they see them. In order to assess the circumstance and classify the occurrence as dangerous, mental processes are engaged. The physiological aspect encompasses the many physical manifestations that arise as a result of the processes taking place. Some of the physical signs that individuals experience when they encounter a snake include perspiration, difficulty breathing, a feeling of suffocation, a racing heart, and a constriction of the chest. People are scared when they exhibit these signs.

Consequently, it is possible to gauge how individuals feel in reaction to certain situations by observing their physiological responses. One name for the motor expression part is expressive conduct, and another is exhibited body language. The majority of it is people's verbal and facial reactions in reaction to what's happening. When individuals respond to or plan to act upon what is happening, they use both facial and voice expressions. Some examples of this behavior include the use of verbal and visual emotions to convey fear, such as when individuals see snakes.

To change their expression, they may lift their upper eyelids, arch their brows, or lower their jaw. At the same time, their vocal expressions change as the pitch rises, the pauses

grow, and the intensity falls.

Consequently, it is possible to identify people's emotions by the analysis of their muscular expressions. Both the conscious and subconscious parts of the brain are linked to the action tendency component. Ensuring the safety of individuals requires a swift response. When activated, motor expressions carry out the desired action as directed by the brain's motor cortex. As an example, when individuals are scared, their subconscious decides to get them out of there as soon as possible. Finally, people's subjective experiences in reaction to external events are associated with the feeling component.



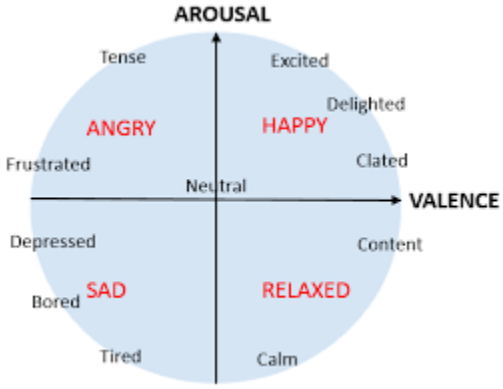
Source: BMC Psychol 12, 503 (2024)

Figure 1.5: Parrott classification of human emotions



Source: https://www.open.edu/openlearn/health-sports-psychology/health/emotions-and-emotional-disorders/content-section-2.2

Figure 1.6: Facial expressions in six universal human emotions



Source: https://www.mdpi.com/1424-8220/23/3/1598

Figure 1.7: Activation and valence dimensions to describe human emotions

This is very intricate problem, affected by a plethora of criteria and causes. The precise setting in which the emotion detection system will function, the characteristics of the target audience, the available resources, and the job at hand are all aspects that fall under this category. Additionally, it is important to think about how effectively the selected modality works with current systems and infrastructure, as well as how reliable, accurate, and

interpretable it is. Ethical concerns, computing efficiency, processing needs, and data collecting ease all add layers of complexity to the process of modality selection. It is crucial to do a complete review and study of these criteria in order to make educated judgments about the best ER modality for a certain assignment.

1.4 EVOLUTION OF HUMAN COMPUTER INTERACTIONS

There are many different activities that fall under the umbrella of Human Computer Interaction (HCI). A few of its many facets include cognitive psychology, arts, ethics, computer graphics, human factors, and cognitive psychology. Enabling people to connect with computers successfully and efficiently, all these disciplines embrace and integrate, complementing each other's ideas and techniques.

Research in human-computer interaction (HCI) focuses on both the two-way street between people and computers. GUIs (graphical user interfaces) that employ windows, menus, and icons, etc. were originally the only focus of HCI with the goal of making systems more compatible. There has been enormous development in HCI in recent years. From manually customizing each memory location using binary switches to the present day, when we have higher-level programming languages, graphical user interfaces, and input devices, the interaction is much smoother. The capacity to converse with a computer as naturally and easily as with other people is still a long way off, however.

The major goal of HCI shifted away from interfaces as the field progressed and the issues with interfaces became more recognized. Computers nowadays aren't only for getting things done; they're also for learning, communicating, explaining, arguing, debating, observing, and designing. These days, improving human-computer interactions via explanations and justifications is HCI's top priority. With this, we can see that new kinds of "indirect" and "natural" interaction are cropping up. Human-computer interaction involves exploring different ways to enter data, such using voice, body language, and facial expressions. For those who aren't very tech-savvy, they may stand in for the more impersonal tools like a computer keyboard and mouse.

When it comes to improving human-computer interaction, one key element is emotions. When people talk to one another, their emotions are a major factor. Olson, Gary & Olson, Judith. (2003) [5]. However, feelings have not been a major factor in HCI up to this point. Emotional intelligence (HCI) is a difficult but potentially useful field to expand into.

1.5 ROLE OF EMOTIONS IN HCI

Emotions are integral to our daily lives and everything that we do. Computers have surpassed people as the primary means of human interaction. On the Internet, people may join virtual communities where they can talk to one other via their computers. Every part of our responses and interactions—with people, with technology, and with technology itself—are impacted by our emotions, sentiments, and moods. In contrast to human-computer interactions, it takes more time for an emotional reaction to be comprehended in a face-to-face setting.

In order to include emotions into human computer interaction, researchers have studied and developed a variety of methodologies and technologies. A number of current developments in human-computer interaction center on the idea of autonomously sensing a user's emotional state and adjusting its actions appropriately. More and more studies are focusing on this area.

For instance, in order to amplify player engagement, entertainment-related apps could use feedback based on emotions. One example is how computer games may adjust their settings based on the player's emotional state. Another area where emotion reading could improve effectiveness in anticipating individuals' future behaviors is surveillance. Thus, the current techniques for detecting and preventing terrorist acts in public spaces may be improved with the use of emotion-driven technologies. The automated emotion reading technique has many potential uses, and these are only a few of them.

Emotional intelligence and sensitivity are not necessities for all computers. It is acceptable to leave certain machines in their rigid tool state as they are helpful for certain tasks. Having computers adapt to their users organically might enhance human-machine interaction in certain instances, but there are other contexts where emotional information is needed to communicate when and how crucial it is to adapt. The literature talks about these systems that can identify user behavior based on different emotional states, making the interaction between humans and machines more natural. The term "Affective Computing" [Picard, 1997] [6] describes this field. This is the domain in which the studies presented in this thesis operate.

1.6 AFFECTIVE COMPUTING

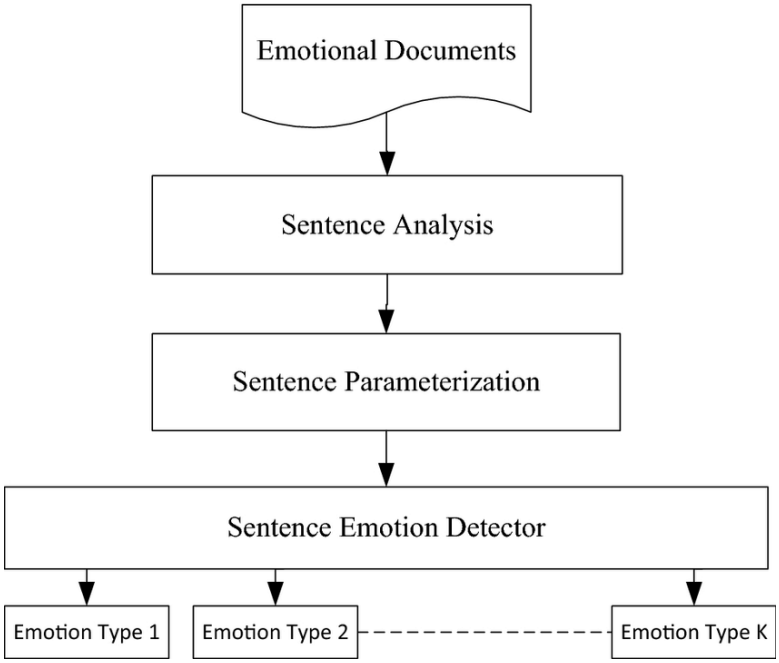
Perceiving, processing, recognizing, interpreting, and simulating human emotions is the goal of "affective computing," a subfield of AI. This study combines elements of computer science, linguistics, neuroscience, and psychology. Systems that are able to identify human feelings are the main focus of human emotion detection, a subfield of affective computing. The three main components are sense, process, and recognize. A schematic depicting the basic building blocks of an emotion identification system for humans. Starting with the sensing module, which records information about facial expressions and actions, the process continues. The data created by humans may be measured by the sensors built into the recording devices. The data that has been collected is sent into the processing module, which then does preprocess to filter out any irrelevant information. By doing the transformation on the valuable data, it is able to retrieve the pertinent details. As part of the supervised learning process, a recognition module takes the extracted data and uses it to train the model. In terms of characteristics, it constructs the model to detect patterns in the extracted data.

Decisions and emotion detection are both handled by the produced model. The acknowledged human emotions have a major part in the highest-level decision-making process. An effective method for human-machine interaction is produced by the relatively young topic of affective computing. This enables the computer agent to have natural-sounding conversations with people. Education, healthcare, gaming, and a host of other possible domains for applied emotional computing have also found use for it.

Research into the incorporation of emotional capacities into computational systems is a relatively young area known as affective computing. Its foundation is on research and hypotheses that highlight the role of emotions in learning, memory, decision making, and almost every other cognitive activity.

Assuming a human talent already exists, it will be possible to replicate it in a machine with enough work and time. To achieve this goal, we must make significant strides in comprehending and using the existing literature on human-to-human interaction [Picard, 2000] [7]. Researchers are tasked with discovering the building blocks of these affective systems and how to integrate them with other capabilities to create systems that are really intelligent, personalized, and captivating.

The term "affective computing" is used to describe computations that are associated with, derived from, or impact emotional states (Picard, 1997) [6]. Machines need to be capable of perceiving human emotions and adjusting their operational behavior accordingly, responding appropriately to certain emotions. This means the system will be easier to use and provide results that are more in line with how humans act. User confidence in emotion-enhanced human computer interfaces is predicted to grow, leading to wider adoption and more practical uses for these interfaces.



***Source*:** *Ghanbari-Adiv, et al. (2019). https://www.researchgate.net/publication/332962981_Text_Emotion_Detection_in_Social_Networks_Using_a_Novel_Ensemble_Classifier_Based_on_Parzen_Tree_Estimator_TPE [8]*

Figure 1.8: Generalized block diagram of automatic human emotion recognition system (Neural Computing and Applications)

1.6.1 Relevant Case Studies and Applications of Affective Computing

Numerous fields may benefit from affective computing. What follows is a list of a few of them. An intelligent web browser might provide more information about the things or subjects the user is interested in until it notices that their interest is waning. An emotional assistant agent might consider your mood and level of activity while intelligently filtering your email or calendar.

Computer agents may improve their teaching abilities by learning the user's emotional state. By observing the student's behavior and preferences, a computer agent may tailor its interactions accordingly. The e-commerce sector, the entertainment industry (including video games, virtual reality, and training programs), and other related fields may reap enormous benefits from this line of work, much as e-learning has. Several studies have looked at the correlation between students' and teachers' rapport and how they act in class. A recent study by Hughes et al. (1999) [9] found that students' aggressive behavior in class decreased when teachers and students had a warm and personal connection. If virtual assistants in the classroom could read students' emotions, they might be a lot more helpful in their tutoring roles.

Because of the requirement to provide customers with realistic, human-like interactions that are mood-aware and responsive, the gaming and entertainment industries are the most apparent commercial users of emotion-sensitive technologies. Similarly, automated customer service and contact centers may be enhanced with emotional sensitivity to identify and address staff and customer annoyances and frustrations. Psychiatric diagnosis, smart toys, and lie detection are among the other uses of automated emotion recognition [Cowie et al., 2001] [10].

Studies aimed at identifying people's emotions based on their interactions in the scientific and professional fields might also be very interesting to cognitive scientists. Security guards, attorneys, and police officers, for instance, pay close attention to nonverbal clues about attitude and other such factors.

Several prominent HCI projects are involved in emotion recognition, including AUBADE (A wearable EMG augmentation system for robust emotional comprehension), SIMILAR (The European Taskforce for the creation of human-machine interfaces SIMILAR to human-human communication), and ERMIS (Emotionally Rich Man Machine Interaction System). There are a number of initiatives in this field at the MIT Media Lab. An example of a social robot that actively seeks out human interaction for the purpose of both communicating and interacting with it is Kismet [Breazeal, 2003; KISMET site] [11]. The robot's ability to respond appropriately and appropriately timed actions depends on the coordination of its emotional and expressive motor systems.

A watershed moment in Japanese entertainment history occurred in 1999 with the introduction of Sony's AIBO [AIBO, online], also known as the Entertainment Robot, an electronic robot. The acronym for "Artificial Intelligence RoBOt" is AIBO. By developing a synthetic entity that closely resembles a biological being, Sony bridges the gap between humans and robots. AIBO's emotions and instincts are hardwired into his brain. His instincts drive him to act in order to satisfy his need. When he's happy, his joy level rises; when he's unhappy or furious, it falls. His development and character are heavily impacted by your reactions to his feelings.

An artificial system for understanding human emotions using data from facial expressions, voice, and body language is called Emotionally Rich Man-machine Intelligent System (ERMIS) [Fragopanagos and Taylor, 2005] [12]. The overall The ERMIS project's objective is to develop a functional model of an HCI system that can detect and respond to user emotions including interest, boredom, and rage.

1.6.2 Potential Concerns

The ability to feel, measure, monitor, convey, influence, and control emotion might be used for destructive or otherwise unwanted reasons, despite the fact that affective technologies permit a broad range of intriguing new and good uses. There is usually a shadow side to technology that makes people rethink social, ethical, and philosophical issues. Sketched below is one such example.

There are a lot of concerns about the benefits and risks of collecting and potentially sharing people's emotional data via sensing, storage, and real-time transmission. Take this scenario: a driver wants her car's navigation system to detect and adapt its voice to her mood, which makes driving safer. But she also wants to keep this information from her insurance company, because they might charge her more if they discover she drives while angry.

How can individuals whose wishes are to have their emotional data felt and interpreted be guaranteed their privacy in this age with such technology? Similarly, from HAL to Terminator, across the entertainment industry, computer systems implanted with all the emotionally antagonistic views against the whole human species have been a recurring motif for decades. It would be highly irritating to expect a computer to continually display its internal state, just as people do not constantly display their emotional moods. For

conveying one's emotional condition, what is the sweet spot, and what elements determine it? Emotional intelligence in computer decision-making is a tricky subject. Finding a happy medium where the benefits may be enjoyed without succumbing to illogical actions is essential.

1.7 DEVELOPING GUIDELINES FOR EXAMINING AUTOMATED EMOTION RECOGNITION'S STATUS AT THIS MOMENT

One of the trickiest parts of thinking about how AI and humans interact is making sure users are a part of it all the time. Facilitating user interaction, information exchange, and feedback, as well as the acceptance of AI suggestions and judgments, are all part of this process. These user-centric concerns are in addition to the technical ones that need addressing, such as the AI technology's maturity level and the platform's or interface's usability. For Human AI systems to be widely used and implemented, it is essential to achieve equilibrium between technical complexity and user involvement.

Numerous strategies employ artificial intelligence (AI) and technology systems to try to address the issues of user engagement and acceptance. The Technology Acceptance Model (TAM) is unique in that it takes into account how useful and simple people believe technology to be as crucial factors in its adoption. Another model that builds on TAM is the UTAUT, which adds social influence, enabling environments, and behavioral intents as additional components. Furthermore, the PSD model stresses the significance of persuasive elements in technological design to impact users' mindsets and actions.

Especially when it comes to artificial intelligence systems, ethical issues play a crucial role in user acceptance models. Traditional models such as TAM, UTAUT, or PSD do not explicitly include these factors, but they are becoming acknowledged as essential components of technology adoption and user engagement. When people rely on AI systems to do urgent or time-sensitive activities, real-time performance becomes an important factor in user acceptance models. Perceived usefulness and ease of use are two characteristics that often overlap with real-time performance, which shapes users' perceptions and experiences. However, classic models like TAM or UTAUT do not explicitly address this intersection.

As one of the aims of this study is to equip us to conduct a current status assessment, we have compiled a list of criteria and provided extensive explanations for each one based on

the models and articles cited. This is why we set out to integrate the social, technological, and scientific approaches—the three cornerstones of emotional recognition—into our work.

None of the Models specifically mention accessibility as a criterion. But it may influence people's opinions and actions toward a system in a roundabout way. In both the TAM and UTAUT models, the concept of "Facilitating Conditions" allows for an indirect approach to accessibility.



Source: https://bogolovan.com/emotional-expression-or-how-open-are-you-in-expressing-your-feelings/

Figure 1.9: Emotional expression.

Data privacy, justice, openness, and account-ability are only a few of the many facets that fall under ethical consideration. When AI systems act ethically, protect user privacy, and explain their decisions clearly, users are more inclined to utilize and trust them. As an added layer of sensitivity, data pertaining to emotion detection is often associated with mentally sensitive material. Data deemed extremely sensitive includes not just the datasets needed to train the models and the modalities from which the data is taken, but also the analysis findings themselves.

Ease of Use refers to how little mental work or stress a system's users report experiencing when interacting with it. This criterion is highly regarded in the TAM model. Users tend to choose AI systems that respond instantly, analyze data quickly, and allow for frictionless interactions in real time. User happiness and adoption may be severely impacted by

delayed reaction times or system performance lags. To ensure users see the system as effective and successful in resolving their demands, real-time performance is a crucial element in the assessment of AI systems.

The chosen modality should be able to be integrated into the multimodal system without major technological hurdles or limitations in terms of resources. One way to measure the credibility and rigor of scientific studies is to look at their publication count. There are a lot of publications in a field, which can be a sign of activity and production, but it's also important to look at the quality, impact, and relevancy of those articles. Nevertheless, this evaluation uses the quantity of publications as a measure of output and the attention that scientists in the area have received.

Historical Context, as we've already seen, there are several ways in which people express their feelings via speech. Accordingly, keep in mind that even with high accuracy rates, depending on just one or two methods to identify emotions may not provide us with all the information we want when it comes to emotion identification. For example, even if we're rather good at identifying those words, we're only picking up on a fraction of the total feelings conveyed, since words only account for 7% of emotional expression (according to the Mehrabian rule) [13.

1.8 MULTIMODAL EMOTION RECOGNITION

A new topic of emotional computing, multimodal emotion detection is only starting to get off the ground. Combinations of auditory, visual, and physiological inputs are known as multimodal signals. The video and audio signals show how people react when they're feeling a certain mood via their facial expressions and words. When a person experiences an emotion, their physiological signals reveal it via changes in their temperature, breathing rate, electrical activity in the brain, and muscular current. In order to differentiate between different human emotions, researchers are working on a multimodal emotion detection system, which is a significant pattern recognition challenge. To improve emotion identification, it incorporates data from people's face expressions, voice, and bodily symptoms. Human facial expressions and audio data are taken into account for multimodal emotion identification in this thesis. Consideration of speech and facial expressions for emotion identification is warranted for a number of reasons.

For starters, there are a lot more places to put research into multimodal emotion

identification using facial expressions and voice. With the advent of web 2.0, data takes on new characteristics. Web 2.0 sites, such as Blogs, social media marketing, social networks (like Facebook, Twitter, and others), and other platforms provide a wealth of multimodal (audio-video) subjective data. Businesses may learn a lot about social market trends by mining multimodal data for information about people's emotions. The multimodal emotion detection system is essential for the audiovisual database's indexing, retrieval, and summarization of speech and facial expression meaning information. Secondly, for real-time applications in domains including security, entertainment, education, and medicine, emotion identification via facial expressions and voice provides a practical and affordable option. Electronic health records, online education, home automation, and other forms of electronic monitoring need its use in smart cities. In order to record human speech and facial expressions, cameras and microphones are required.

The microphone and camera are inexpensive and undetectable by humans. As a result, they provide human emotion identification using speech content and impromptu facial gestures. The last argument is that identifying emotions in speech and facial expressions presents several challenging research challenges. When people sing, laugh, weep, or speak, their vocal tracts and facial expressions change in distinctive ways. There has been no evidence of any resemblance between their faces in terms of hair color, nose shape or size, cheek size, eye color, eyebrow shape or size, mouth shape or size, or chin color. Additionally, their vocal speech has varied; adult males typically use a frequency range of 85–180 Hz, while adult females use 165–255 Hz. As individuals age, their facial features and vocal qualities undergo gradual transformations. Building a generic multimodal system for emotion identification is tough due to the variety of speech and facial emotions. The problem of multimodal emotion recognition requires a reliable and efficient computing solution.

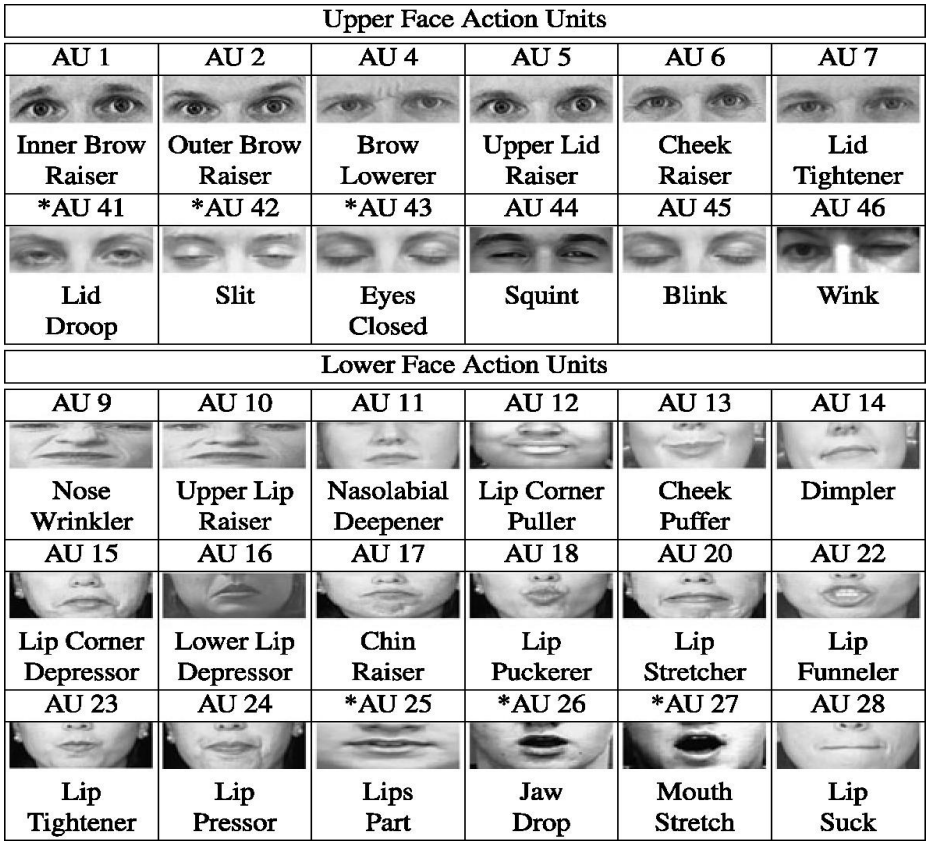
1.8.1 The role of multiple modalities

You can learn a lot about people's emotions just by looking at their faces and listening to their voices. To communicate, people use both words and facial expressions at the same time. The "7%-38%-55%" guideline is how they evaluate other people's actions. Words make up 7% of their focus, whereas speaker tone and facial expressions account for 38% and 55%, respectively. Communicating ideas and emotions, people employ both oral and facial clues at the same time. In most cases, when individuals react to something, their

subconscious minds come up with the voice and facial reactions. To reliably measure human emotions, these constructs are evaluated.

There is a wealth of information about human behavior conveyed by the movement of facial muscles. When the muscles of the face are moved, the face takes on a different look. An industry standard known as the FACS (Face Action Coding System) specifies and codes the motions of the face muscles. Figure 1.10 shows the Facial Action Units (FAUs) that were used to describe the motions of the human face muscles. When there is no expression or movement in the face, the facial action unit is labeled 0. The FAUs are called after the many facial muscles that move in different parts of the face. Raising the inner brow (AU 1) and the outer brow (AU 2) are the names of the frontal forehead muscles that are responsible for this action.

The AU 4 is the name of the procerus muscle's depressor (brow lower). For the elevation in the superior tarsal muscle, the upper lid raise (AU 5) is specified. Other FAUs have also shown that face muscles can move. Important information on human emotional state may be gleaned from the existence of facial action units. A happy expression is indicated by the cheek lifting (AU 6) and corner tugging of the lips (AU 12), according to the psychiatric research. Smiling faces indicate that individuals are experiencing happiness. Raising the inner eyebrows (AU 1) or lowering them (AU 4), they show their sadness. As a sign of remorse, they have bent their chin (AU 17) and pushed the corners of their mouths (AU 15). When they were sad, they would sometimes separate their lips in different ways (AU 25) and stretch them (AU 20). Wrinkles on the nose are a way they display their displeasure (AU 9). For AU 4, they drooped their eyebrows, and for AU 5, they would sometimes lift their upper lids. To convey wrath, they would lower their jaw (AU 26), lengthen their lips (AU 20), or change the shape of their lips (AU 25). Raising the three brows—the upper (AU5), inner (AU 1), and outer (AU 2)—communicates dread. Fearful facial expressions might also be seen on occasion with lowered eyebrows (AU 4), dropped jaws (AU 26), and extended lip corners (AU 20). Facial expression-based emotion identification relies heavily on the facial action unit profiles.



Source: https://library.fiveable.me/neuromarketing/unit-6/facial-coding/study-guide/qSKHJCKyaDPlQfZ5

Figure 1.10: Facial muscles movements defined in Facial Action Coding System

The human respiratory system, voice chords, and articulators work together to produce speech. In order for the vocal cords to produce sound, the lungs must pump air into them, which in turn creates air pressure. Pulses of audible vibration are produced when the vocal cords are subjected to air force. It is the interplay between auditory pulses and articulators (tongue, cheeks, lips, etc.) [14] that produces the complexity of voice. Considerable data on human conduct is conveyed by the voice in speaking. It conveys both the speaker's emotional state and the meaning of what they are saying. A speaker's tone depicts the qualities of their voice that are conveyed via speech, including pitch, intensity (loudness or vigor), quality, pace of speech, and attitude.

The message's wording reveals the speaker's general ideology. You can tell a person's emotional state just by listening to their tone and the polarity of their words. When angry, people tend to talk with a high pitched, intense, breathy voice, somewhat faster than normal

pace of speech, and an abundance of negative polarity phrases. When they're happy, they utter more words with positive polarity, utilize a higher or lower vocal pitch, and sound intense, gasping, or shouting. In addition, when individuals are unhappy, they tend to talk more slowly, with a lower tone, and less volume. They achieve a rich, deep, and crystal-clear sound. When they're upset, they could express it using more negative polarity phrases. When individuals are scared, they talk faster and higher in pitch. Even with irregular speaking, a loud voice is par for the course. Typically, when people want to express dread, they employ more words with negative polarity.

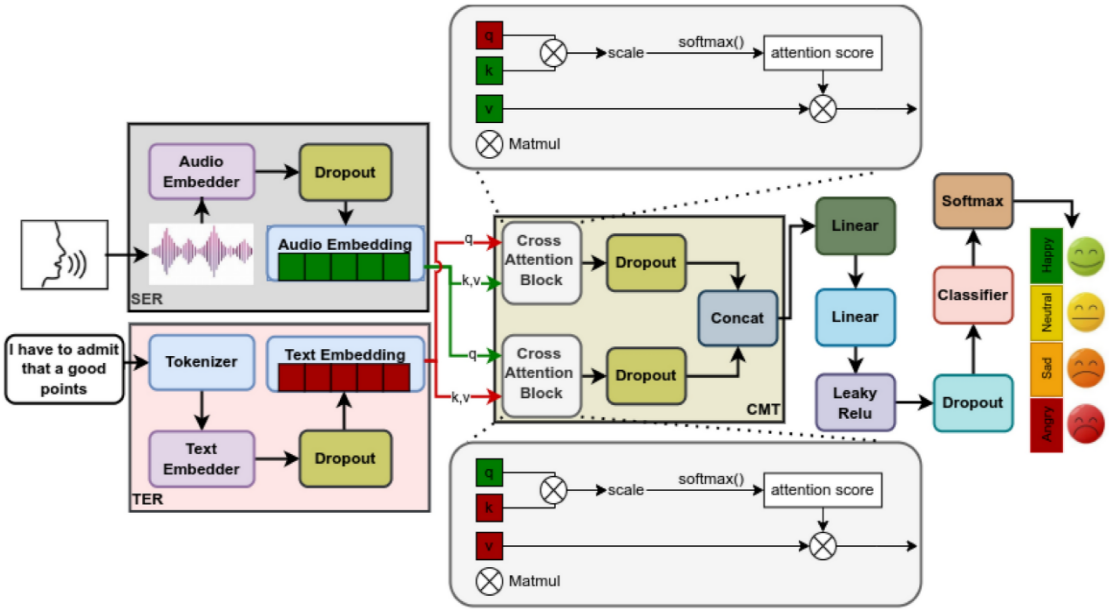
1.8.2 Current approach

The three most cutting-edge approaches to emotion detection in affective computing are as follows: (a) a unimodal approach based on images and videos; (b) a multimodal approach based on audio and language; and (c) a combination of the two. Emotion recognition systems have been the subject of intensive research into ways to increase accuracy while decreasing processing cost. Face expressions are the only determinants of human emotion in the image/video unimodal approach. Humans' auditory systems have not been taken into account.

The audio/linguistic unimodal method, in contrast, has only been able to identify human emotions via speech. Human face emotions and visual information have not been taken into account. But facial gestures and words are the main ways that individuals express themselves. In order to express themselves, they could employ all of these structures at once or choose one. As a result, we have been studying how to recognize emotions in facial expressions and voice by fusing visual and aural modalities. The four primary components of a multimodal emotion detection technique are a multimodal dataset, preprocessing, feature extraction, fusion, and classification. A generic pipeline for multimodal emotion recognition utilizing speech and face recognition is shown in Figure 1.11 expressions.

The features and scene contents taken into account in the audio-video clips of multimodal emotion datasets operationally determine the emotion recognition performances of the most advanced methods. Human voice and facial emotions are among the many audio-video emotion datasets that are now available. Most multimodal emotion datasets consist of unilingual speech and facial expressions. The individuals are videotaped in a controlled

setting with the right lighting effects, away from background noise. No objects in the backdrop move, and the subject's location is controlled. In a controlled laboratory setting, the subject and the acquisition equipment (camera and microphone) are kept at a constant distance. The study of speech emotion recognition and facial expressions across languages is still in its early stages. The SEWA dataset contains the facial expressions and multilingual voices of people who saw product advertisements.



Source: https://www.nature.com/articles/s41598-025-89202-x

Figure 1.11: General pipeline for multimodal emotion recognition using speech and facial expressions

The performer's ethnicities included Chinese, British, German, Hungarian, Greek, and Serbian. Multilingual English and Turkish audio-video clips TV episodes and movies are included in the BAUM-2 collection. Metrics for accurate emotion identification have been impacted by the level of complexity in both spoken language and facial expressions. Actors' and actors' voice and facial expressions are the most common experimental validation data for emotion detection. In order to learn how to express themselves more realistically, some researchers have turned to more spontaneous methods of learning facial and vocal expressions. When it comes to identifying emotions, the level of complexity in spontaneous facial and verbal responses is higher. Accuracy rates for the emotion recognition approach were 95% on the acting RML dataset, 84% on the induced

eNTERFACE dataset, and 55% on the spontaneous AFEW dataset. Increasing the complexity of voice and facial emotions brings them closer to real expressions, but it also decreases the accuracy of emotion identification algorithms.

Important information is extracted from the audio and video modalities during the preprocessing phase. Included in this are methods for detecting and tracking faces, identifying regions that include voice or non-voice, and selecting speech segments and peak stages of facial emotions. Within computer vision, Numerous studies have been conducted on the recognition of human faces and tracking in video frames.

Robust real-time automated face recognition and tracking systems often use the cascade face detection approach. Signal processing has also done much research on audio region identification for both voice and non-vocal content. The speech and non-voice parts of audio signals have been identified using temporal and frequency domain characteristics. As a means of communication, humans use their face muscles and vocal tract to convey a range of emotions. During the peak period, individuals show the most expressive range of motion in their face muscles and voice tract.

The most useful information for emotion detection comes from the peak stage frames of audio-video signals, which include both facial emotions and voice. When it comes to multimodal emotion identification, the peak stage selection problem has received very little attention. First, dissimilarity analysis; second, cluster analysis; and third, divergence analysis are the three main kinds of existing peak stage selection methodologies. According to the dissimilarity analysis technique, the more distant from other frames the contents of the peak frame are. Every frame of the movie has its Local Phase Quantization (LPQ) characteristics calculated using the Maximum Dissimilarity (MAX-DIST). The N × N chi-square dissimilarity matrix of N visual frames is constructed by it. The six frames are selected as the ones with the greatest dissimilarity scores peak stage frames. The emotion recognition accuracy on video clips from the enter FACE dataset was 37.22% and, on the BAUM, -1a dataset it was 46.60% using the multiclass Support Vector Machine (SVM) classifier using linear kernel function.

The video's important frames have been located using the Minimum Sparse Reconstruction (MSR) technique. In order to synthesize the visual frames, it calculates the least square

reconstruction error. On the AFEW (Acted Facial Expressions in Wild) dataset, using video clips, the multiclass SVM with chi-square kernel function classifier achieved an emotion recognition accuracy of 49.9%. The accuracy of the multiclass SVM classifier was enhanced by 10.58% compared to the video baseline when the MSR approach was combined with the audio feature descriptor. The visual frames have been sorted into n separate groups using cluster analysis techniques.

There are visual frames of the same sort in each cluster, and there is very little space between them. To summarize the contents of the video stream, they used the landmark coordinates of facial expressions and the k-mean clustering algorithm with k=4. We provide the Dendrogram Clustering (DEND- CLUSTER) technique for human emotion identification peak frame selection. A pair-wise distance matrix of LPQ features from N frames was calculated, with dimensions N × N. The visual frames were grouped into six groups using the agglomerative link clustering approach. On the eNTERFACE dataset, the multiclass SVM with linear kernel function classifier attained an accuracy of 40%; on the BAUM-1a dataset, it was 55.70%.

The distance in geometric units between the visual frames and the modified neutral face subspace is determined using the divergence analysis approach. In order to determine peak stage frames, the Emotion Intensity (EIFS) technique has relied on neutral facial expression frames. It built a changed neutral subspace by applying KL transformation on neutral frames in the Cohn-Kanade dataset. By calculating the smallest possible distance from the neutral subspace to the video's neutral face frame, it finds the neutral frame. When choosing a neutral face frame, it looks at the top six frames and uses their greatest distance from the chosen frame as a metric for peak stage frames. The emotion detection accuracy on the eNTERFACE dataset was 30.38% and, on the BAUM, -1a dataset it was 52.06% using the multiclass SVM with linear kernel function classifier.

The effective detection of human emotions in audio-video data relies on feature extraction. The audio-video signals are transformed into the feature vector that the classification model uses via the methods for feature extraction. The feature extraction procedure from audio and video signals into a feature vector has been the subject of several presentations throughout the years. Facial expressions in videos have been represented by models that use geometry and texture feature analysis. Here, we introduce the Constrained Local

Models (CLMs) that can identify and follow face locations in video sequences. When it comes to detecting and tracking facial point landmarks, The CLNF concept, or Conditional Local Neutral Field outperforms other techniques. The techniques used for analyzing texture features are also known as appearance feature methods. In relation to the change in expressions on the face, they calculate the gradient change. For the purpose of representing face expressions, the Haar features and the Gabor wavelet are used. In order to characterize the face area in picture frames statistically, using tools like histograms, there are statistical texture features.

The face feature vector is calculated using the Histograms of Oriented Gradients (HOG) and the Pyramids of Histograms of Oriented Gradients in Three Orthogonal Planes (PHOG-TOP). They made use of image-based edge gradient computation to feed into a gradient histogram. Facial expressions are better recognized when the Local Threshold Binary Pattern (LTBP) approach along with the Histogram of Oriented Gradient (HOG) technique are combined. Emotion detection in low-resolution photos is made possible by the Pyramidal Local Binary Pattern (PLBP) algorithm. The accuracy of emotion recognition by SVM classifiers is improved by PLBP features when compared to PHOG, Gabor, and LBP features.

When it comes to emotion categorization, we compare the Local Phase Quantization (LPQ) characteristics in conjunction with the Local Binary Pattern (LBP). LPQ features have improved the accuracy of the linear SVM classifier in comparison to LBP features. Local Binary Patterns on Three Orthogonal Planes (LBP-TOP), Volume Local Binary Pattern (VLBP), and Local Phase Quantization on Three Orthogonal Planes (LPQ-TOP) methods are used to calculate the spatio-temporal texture aspects of face expressions. Verbal information, such as how people express themselves via language and intonation, is also a part of the audio signal.

Different methods for extracting speech There are four types of features: time-frequency (spectrogram) domain, frequency-domain, and time-domain., or linguistic domain. The time domain characteristics are calculated straight from the audio signal samples, without any processing applied. In order to classify emotions, the time domain properties of energy and zero crossing rate are calculated. To calculate the frequency domain characteristics, one must first estimate the distribution of energy in the voice signal's frequency

components using the Discrete Fourier Transform (DFT). Prosodic, spectral, cepstral, and ad hoc features are the four primary types of frequency domain features. The speech signal is measured and described in various ways by these groups. One common use of prosodic characteristics is the categorization of speech emotions. Pitch, volume, formant, duration, and vocal quality are all prosodic characteristics.

Over the range of audible frequencies, the spectral properties measure the dispersion of spectral energy. In order to make the voice signal more comprehensible to humans, cepstral approaches are used. There is a difference between cepstral and perceptual characteristics. Perceptual Linear Prediction (PLP), Human Factor Cepstral Coefficients (HFCC), Bark Frequency Cepstral Coefficients (BFCC), Relative Spectral-Perceptual Linear Prediction (RASTA-PLP), Mel Frequency Cepstral Coefficients (MFCCs), Revised Perceptual Linear Prediction (RPLP), and Gamma Tone Cepstral Coefficients (GTCCs) are some of the notable perceptual feature extraction methods. The voice signal is represented as spectrograms using the time-frequency analysis techniques. In a nutshell, it gives you the time, frequency, and decibel (DB) value of each frequency component of the voice signals. The area of natural language processing known as linguistic-based emotional content analysis is expanding at a fast pace. The voice signal was transformed into language words by them. A number of characteristics may be retrieved from text using different methods; they include key spotting, an ontology-based method, and a statistical method.

To correctly identify human emotions using audio and visual data, the penultimate stage is fusion and classification. There are three main types various multimodal fusion techniques, including decision, model, and feature fusion. This method of merging audio and video information prior to feeding them into an emotion classification model is known as an early fusion strategy. At the first step, feature fusion takes use of the association between audio and visual modalities to enhance the method's accuracy in emotion identification. However, feature fusion has a major flaw in that it cannot account for missing data from a single modality. Simultaneous emotions on the face and in the mouth are difficult to depict. Before or after speaking, people display facial emotions. There is a time lag between facial emotions and spoken words in audio-visual communications. In addition, "intermediate fusion" describes the multimodal fusion step of the paradigm. In order to classify data, the model looks at how speech and facial expression characteristics are correlated with one

another. It enables the simultaneous analysis of audio and video in order to leverage the related facial expressions and significant relational information included in spoken words. Multimodal emotion categorization makes use of a number of different model stage fusion algorithms.

A few examples of these models are the triple-stream dynamic Bayesian network model, the BLSTM neural networks, the EWSC-HMM, the Hidden Markov Models (HMMs) [15], and the 2H-SC-HMM, which are based on hierarchical alignment. In order to identify human emotions, these algorithms understand the pattern of time sequences in video and audio modalities. A late fusion method is also used when multimodal fusion occurs during the decision-stage. Expressions on the face and in spoken language are taught to use different emotion categorization models. It combines the probability of two separate emotion categorization models for audio and video into a single output. Multiple approaches exist for merging classifiers' output class probabilities.

1.8.3 Research Issues and Challenges

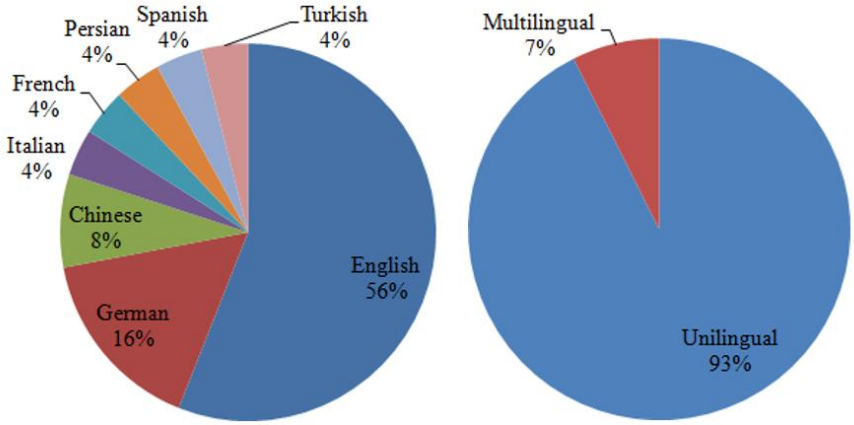
The field of research aiming to bridge the gap between people and robots is making great strides. Research into emotion computing has made great strides over the last several years. But there are still several problems and obstacles that computers using AI approaches face when trying to accurately identify people's emotions from things like facial expressions and voice. What follows is a description of the issues and challenges related to multimodal recognition of human emotions.

- **The Indian context does not have a standardized multimodal emotion dataset.**

The bulk of current One noteworthy and consistent conclusion from the literature study is that multimodal emotion databases contain the unilingual speech and facial expressions of foreigners. The language demographics from the multimodal emotion datasets are shown in Figure 1.12. The majority of earlier multi-modal emotion datasets recorded respondents' voice and facial emotions in their native tongues. 56% of datasets consist of the voice and facial expressions of subjects in English. English has a much greater multimodal emotion dataset than other languages proportionate score. These languages include German, Turkish, Persian, Italian, Chinese, and French.

Researchers discovered that the multilingual speech and facial expressions of immigrants

were only included in 7% of earlier emotion datasets. Among them were Chinese, English, German, Greek, Hungarian, and Serbian languages spoken by the actors in the SEWA dataset. In their original tongues, the performers convey their emotions. There are Turkish and English facial expressions and multilingual speech in the dataset BAUM-2. The clips of audio and video were taken from various television series and movies. The performers failed to convey their feelings in a variety of languages. Using one's native language to convey one's emotions is quite normal. In India, the 2011 census found that the following languages were spoken as first languages: Bengali (8.30%), Urdu (4.34%), Marathi (7.09%), Telugu (6.93%), Tamil (5.89%), Gujarati (4.74%), Kannada (3.73%), Odia (3.20%), Malayalam (2.97%), Hindi (43.63%), Punjabi (2.83%), English (0.02%), and Sanskrit (0.00185%). Speaking the foreign language is a way for some individuals to express themselves. According to the research, no standardized multimodal emotion dataset exists for the Indian setting that records the facial expressions and voice of individuals who speak more than one language. [16]



Source: https://www.researchgate.net/figure/Demographics-of-languages-in-existing-multimodal-emotion-datasets_fig1_365473756

(a) Statistical distribution of unilin-spoken languages (b) Statistical distribution of gual speech and expressions among monolingual and multilingual speakers collection of emotions across several domains

Figure 1.12: Language distribution in current multimodal emotion datasets.

- **Restrictions on the range of emotions that may be conveyed by Speech patterns in many languages and facial expressions**

The diversity of words people employ to convey their feelings is known as their range of speech. Participants' facial expressions and multilingual speech from product ad conversations are available in the SEWA dataset. Multilingual audio-visual clips taken from English and Turkish TV shows and movies make up the BAUM-2 dataset. The SEWA and BAUM-2 databases do not include respondents who speak more than one language.

All of the participants are fluent in their native tongues and hail from diverse cultural backgrounds. People have expressed a wide range of emotions using a variety of words in previous multilingual multimodal emotion datasets. "I won; can you believe that!" one male respondent in the BAUM dataset said. In a state of joy, I am wealthy now. He demanded, "Return it to me!" i'm angry and want to know whose property you're stealing from. The analysis reveals that the individuals' words change depending on their emotional state. When speakers use a variety of words to convey their emotions, the system has an easier time picking up on their tone of voice. The algorithm picks up on the terms "won" and "rich" when people are joyful and "stealing" when they are angry. It categorizes words according to whether or not they include certain frequency components. When subjects use the same words in multiple phrases to reveal different emotions, the system's choices become unreliable. As an example, the person may have stated "open the window of my room" while feeling glad or angry. Speakers' tone-sensitive information reduces the system's accuracy. In cases when participants have used identical words in phrases to communicate distinct emotions, it becomes tough for the algorithm to determine people's moods.

- **Limited effort to evaluate reliability of annotated clips in the multimodal emotion datasets**

Human voice and face expressions captured in audio-video recordings, as well as

annotators' subjective evaluations of these clips, make up the multimodal datasets. You may trust Due to the validity and reliability of the annotator responses; multimodal emotion recognition systems are trained and tested using the recorded audio-video clips. An essential component of reliability analysis is evaluating the replies of both intra- and inter-raters. The degree of agreement between the annotators' assessments of the audio-video clips is shown by the inter-rater reliability study. The degree to which several annotators' comments on the same audio/video clip are similar is determined by this. Analyzing how well annotators agree with one another when asked to assess the same audio/video file at various times is what intra-rater consistency analysis is all about.

Reliability study on annotator replies to measure subjective assessment quality has been undertaken by very few studies. But most studies haven't checked how consistent annotators' answers are when they subjectively rate recorded audio-video snippets. Anger and disgust are two human emotions that are expressed via facial expressions and voice more often than others. Subjects' speech tones and the profile of their active-passive facial action units exhibit striking similarities when expressing disgust and fury. Because of this, it becomes difficult for the annotator to classify the overlapping expressions in audio-video recordings as emotions. To evaluate the degree to which the annotators' judgments align with one another and with one another's, the consistency analysis includes crucial components. It gives the subjective evaluation of the recorded clips' quality. It proves that datasets for multimodal emotion identification are reliable and easy to use.

- **Inadequate experimentation on multimodal feature for emotion recognition in multilingual tone-sensitive speech and facial expressions**

Feature extraction techniques take video and audio information and convert them into a feature vector that the classification model can use to understand the speaker's emotions and body language. Speech signal processing and computer vision have seen a proliferation of feature extraction techniques for both audio and video over the years. Bio metrics, medical image processing, face analysis, motion analysis, and image quality analysis are just a few of the many computer vision application areas that may benefit from the models developed for geometry and texture feature analysis.

A number of Face position detection and tracking in photos has previously been studied

using Constrained Local Models (CLMs). In terms of facial point landmark recognition and tracking, the Conditional Local Neutral Field (CLNF) model performs better than other methods. Furthermore, models for textures analysis that use the Local Binary Pattern (LBP) technique include quite popular since they are easy to use, have a strong discriminative capability, and need fewer computing resources. To solve texture classification difficulties, one may use the LBP or one of its derivatives, such as the Complete LBP, the Weighted Local Binary Pattern (WLBP), or the Local Binary Pattern (CLBP) with Three Orthogonal Planes (TOP).

Application specificity is a hallmark of the characteristics retrieved using LBP and its variations. Finding the best LBP and variant-based feature set to interpret facial expressions for accurate emotion classification becomes an important area of study. The approaches for extracting speech features may be roughly grouped separated into three categories: frequency domain, time-domain, and time-frequency domain. Features in the time domain analysis is a mathematical technique that uses the voice signal to directly calculate measurements with regard to time. Energy, energy entropy, and zero crossing rate are time domain speech properties. Instead of focusing on time, frequency domain feature analysis takes frequency components into account. The audio stream is transformed into frequency bands and magnitude data using the Discrete Fourier Transform (DFT). Mel-frequency cepstral coefficients (MFCCs), gamma-tone cepstral coefficients (GTCCs), spectrum characteristics, and pitch are all properties that exist in the frequency domain. In order to analyze the signal's amplitude, frequency, and time component, time-frequency domain analysis converts the spoken signals into spectrograms. The field of voice feature analysis for emotion categorization across languages has not seen much experimental activity. When dealing with tone-sensitive multilingual speech, finding the best collection of speech features to use for emotion detection becomes more difficult. Hence, to find the multimodal feature set for emotion identification in tone-sensitive speech and facial expressions, it is necessary to conduct experimental comparisons of different approaches for extracting features from audio and video.

The most pertinent and generalizable data for emotion detection is not being calculated because of insufficient effort to analyze and manage characteristics of speech and facial expressions that vary by individual.

While speaking, singing, laughing, or crying, a person's vocal tract and facial expressions take on distinct patterns. Color, form, size, hair color, eyebrows, nose, cheekbones, lips, and chin are only few of the facial profile elements that have not matched. Adult men use a voice-frequency range of 85 to 180 Hz, whereas adult females use a range of 165 to 255 Hz. Their speaking patterns are likewise distinct.

As individuals age, their facial features and vocal qualities change. Developing a generic system is tough because to the variety of speech and facial emotions, both within and across languages. There have been attempts by researchers to use face identification technologies to identify the area of interest in videos. In order to extract the valuable speech data from the audio stream, they have used a quiet removal strategy. To aid with emotion identification, they have built a feature vector that specifies important information about people's audio and video recordings of face expressions and voice. However, they haven't considered the possibility that people's voice and face profile characteristics may differ significantly from one another in audio and video feature descriptions. Developing a general feature description for audio and video that characterizes just the substantial change in speech and facial emotions while ignoring the person-dependent information becomes tough. A more accurate method for recognizing emotions should be possible with the help of the general feature description for audio and video. It will lessen the impact of gender and other characteristics that make each person distinct on the diversity of voices and faces.

1.9 FORENSIC PSYCHOLOGY

A living thing's behavior is its style of acting. Conduct that goes against the accepted standards, principles, and regulations of a community is known as criminal behavior. In our culture, this happens every day. Criminal activity among the general public is a staple of most news broadcasts. Crime reports in Africa's news study and other media outlets are becoming routine. This is also true in other regions of the globe. Images of crime scenes are commonplace on the web. Ads for novels about insurgency are also prevalent. Crime has been a part of the news ever since the earth came into being.

Consequently, we must ask: What motivates individuals to break the law? Is there a way to get crime under control? A branch of psychology called Forensic Psychology (FP) emerged out of the necessity to address these inquiries. Philosophers have long sought to elucidate the motivations behind criminal and antisocial conduct. Criminal law structures

and appropriate punishments for each offense were likewise difficult for jurisprudence experts to decipher. The field of psychology gained widespread attention in the early nineteenth century as a means to better understand and account for criminal behavior. Nonetheless, FP remained underappreciated as a distinct field until a growing number of clinical forensic psychologists started offering compelling evidence in legal proceedings (DeMatteo, 2015) [17].

Experts in FP may now testify on a variety of legal issues, such as whether a suspect's criminal behavior was unintentional or deliberate, and whether or not a defendant has the mental capacity to stand trial in a court of law. Experts in criminal investigation and law enforcement are two roles that forensic psychologists may play (DeMatteo, 2015) [14]. As a result of FP's rapid expansion and rising popularity in fields like the legal system, corrections, the military, and paramilitary organizations, other professionals in the helping professions, including Computer scientists, criminologists, social workers, and lawyers are starting to understand the value of incorporating psychological knowledge into their practice (APA, 2013) [18].

1.9.1 Applications of Forensic Psychology

Because of its pervasiveness and practicality, forensic psychology has maintained some characteristics that make it a complex field. The various important roles that forensic psychologists play in modern society highlight the importance of this profession. They do research that the legal system may benefit from and teach at educational institutions. Studies on the subject of eyewitness evidence, the rights of people, and the decision-making process of juries are among them [18].

Further, they provide direction, assistance, and execution in the following areas: mental health courts, drug courts, family courts, and security consulting services [18].

Government agencies like Nigeria's National Agency for Prohibition of Trafficking in Persons (NAPTIP) and National Orientation Agency (NOA), as well as child and women protection agencies, can seek the advice of clinical forensic psychologists (Mayer, 2002) [19]. Victims of forensic mental health issues, domestic abuse, or deportation may also benefit from the assistance of FP specialists, who can help ease any psychological difficulties they may be experiencing. Ward (2013) [20] states that FP is based on five main areas. What follows is an outline of the main activities things forensic psychologists

undertake in these five key areas, sometimes in combination with other forensic specialists.

1. Psychology of Crime and Delinquency

1. They determine if preschool intervention programs that aim to reduce aggressive behavior in adolescents are effective.
2. Their studies focus on how antisocial conduct emerges.
3. As research policy advisors, they advise lawmakers and government agencies on how to combat devastating criminal behaviors such cybercrime, abuse, corruption, trafficking, and stalking.
4. They work with school staff to identify troubled youths who pose a threat to themselves and others.
5. They create psychological evaluation instruments to study potential dangers faced by youngsters with mental illness.

2. Law Enforcement Psychology

1. They help police officers Figure out the best way to divide up their shifts.
2. They aid employers in the police force in creating psychological profiles of repeat offenders.
3. As part of the hiring process, they create valid and accurate evaluation tools for use by different units and departments to hire duty officers for law enforcement agencies.
4. They educate law enforcement on how to accommodate those with mental illness.
5. Following a conflict or shooting, they provide counseling services to police who may be dealing with issues include suicide, PTSD, depression, and other mental health conditions.
6. Specifically, they create and expand programs to aid the families of police officers, soldiers, and victims of conflict.

3. Victimology and Victim Services

1. Victims and witnesses to crimes are some of the people they assess and treat.
2. They conduct psychological evaluations in cases involving injuries sustained in accidents involving motor vehicles, defective products, sexual harassment, discrimination, medical mistakes, and worker's compensation.
3. PTSD and other psychological responses to being a victim of crime are topics they teach and educate victim support workers on.
4. Those who provide death notification services are evaluated, comforted, and advised by them.
5. They inform social workers on the impact of multiculturalism on victims' access to mental health and further assistance.

4. Legal psychology

1. They analyze the likelihood of child abuse, visitation conflicts, and custody disputes.
2. By doing community research and other forms of inquiry, they aid lawyers in the jury selection process.
3. Their job is to determine if a defendant is competent and worthy to face trial.
4. When making custody judgments, resolving conflicts, or evaluating the efficacy of evaluation processes for different mental health issues, they confer with lawyers and the courts.
5. In both civil and criminal cases, they assess the ability of potential witnesses.

5. Correctional Psychology

1. They provide trustworthy and legitimate evaluation tools for prison staff.
2. Before admitting a prisoner, they determine whether or not the prison's mental health services are adequate.

3. They devise therapy programs for prisoners, both individually and in groups.

4. Reconciling victims and offenders, programs for sex offenders, health education initiatives, and other similar initiatives are evaluated for their efficacy, whether they are targeting juvenile or adult offenders.

5. They offer programs to help correctional workers manage stress.

1.9.2 Challenges of Investigations in Forensic Psychology

Effective and efficient services in the investigation arena need specific approaches and procedures for social forensic psychologists. The credibility of the investigation's findings, however, might be impacted by the prejudices of the specialists questioned. Also, in an effort to escape punishment and be rescued, a suspect presented to a forensic investigator for questioning may adopt a hostile, suspicious, or uncooperative general cognitive "mind set" (Gudjonsson, 2004) [21]. A suspect (interviewee) may act dishonestly, cleverly, pretentiously, with a stressed-out posture, agitated emotions, and even plead for sympathy if they adopt this mentality. This has a detrimental effect on the interrogator's ability to deal with inconsistencies and deceit during questioning, as it might affect the interviewer's perception of the suspect's actual confession. There are two ways of responding to stressful situations, says Gudjonsson (2004) [21]. The Realistic and Logical Approach vs. the Passive and Powerless Position.

- **Logical and Realistic Approach**: The interviewer's motivation to keep going could be dampened when the interviewee actively resists opening up, even though this is an attempt to address the incident or issue at hand. When this occurs, it becomes very difficult, if not impossible, for the interrogator to get the suspect to come clean or confess.

- **A Helpless and Passive Position:** This may make the interviewee more vulnerable as they want to avoid getting into a Figurant with the forensic psychologist (the interrogator), which might alleviate some of the tension.

Saul (2008) [22] claims that suspects who are resistive to confession are more prone to maintain their own truth narrative and reject attempts to persuade them to tell the truth. Forceful treatment, according to some experts, may get suspects to confess, while coercion can change their minds (about telling the truth). Some people may even lie to the

investigator to avoid suffering or penalty by agreeing with the investigation. According to Redlich (2018) [23], there are two perspectives on how an interrogator might coerce a suspect into giving a false confession.

The suspect knows that his or her confession conflicts with his or her private, inside knowledge of the occurrence or scenario, yet he or she may nonetheless cave to the interrogator's pressure. A coerced-compliant fake confession is the name given to this technique. Even when the interrogator knows the agreement is not true, the coerced suspect may be able to handle the pressure of the questioning and eventually give in to the interviewer. Incorrect confessions occur when a suspect is forced into complying. The interviewee may repeat this behavior for many factors, including as but not limited to: trying to win over the investigator, avoiding additional questioning and detention, avoiding actual or perceived physical punishment, or developing a rapport with the investigator that could lead to a genuine confession through reinforcement (Vennard, 1984) [24].

The suspect's internal narrative may help bring about adjustments in some cases. That is, in accordance with what the investigator anticipates, so that the suspect concurs, either publicly or privately, with the interrogator's or interviewer's description of what happened.

The term "coerced-internalized false confession" might describe this situation. Crucially, the suspect may start to believe that the interrogator's account of events is more accurate than his or her own, which is a major problem with this kind of confession. A distinction between this kind of confession and the fabrication of false memories has been proposed by Kassin (2013) [25]. On the other hand, there are variations in the forms of cognitive dissonance (when an individual changes their attitude to be more consistent in their behavior) and compliance that occurs in accordance with authority.
Now that we know that forensic specialists may face problems when using conventional investigation techniques, we need to learn how to conduct genuine inquiries that may be used in court, as well as how to evaluate and advise authorities and clients. Now let's talk about FP inquiry approaches.

1.9.3 Forensic Psychology Investigation Methods

The majority of FP evidence or clues (useful conceptions) are encoded or labeled when they pertain to criminal investigations. What this means is that the reliability and completeness of the remembrance are enhanced by the witness, who is believed to have helped make the memories more accessible (Fisher & Geiselman, 1994) [26]. Forensic psychologists use the following methods to help interviewees or suspects recover from incidents quickly and adequately (Vrij, Leal, Mann, Warmelink, Granhag & Fisher, 2010) [27]:

- **Inverse Sequence**

This manner, the witness might be requested to start explaining the scenario or occurrence from several angles. For example, they can start in the middle, or they can start at the end and work their way backwards to the beginning. One possible question to ask a suspect is when and where they were arrested. From that point on, details like when and how the person was taken into police custody, the nature of the offense, etc.

- **Shift Point of View**

Here, investigators want eyewitnesses to paint a detailed picture of what happened at the crime site as if they were a bystander or a victim.

In order to get a witness to recollect or reprocess the information they've already remembered, investigators utilize two techniques: reverse order and shift viewpoint. Increasing the total number of recall points should help with retrieval of previous and actual events or situations, every journey, supposing that recollections of incidents are kept as networks of connections (Fisher & Geiselman, 1994) [26].
Forensic interviewing involves a broader mode of operation for the cognitive pattern, which includes concerns like the value of a branded, imagined event and a larger variety of specific interrogative tactics (Fisher & Geiselman, 1994) [26]. There have been several additions to studies that have been carried out in both laboratory and field settings since the late 1800s. The degree to which cognitive interviews elicit more honest admissions than other forms of questioning is shown by the weight of the evidence, as stated by Milne and Bull (2003) [28]. In addition to having widespread support among law enforcement, new research has shown that these strategies work well when used with young witnesses.

- **Revealing Deceit and Lie**

Being cognizant of how simple it is to produce a false or misleading confession is fundamental to forensic psychology. In order to determine if interviewees are telling the truth, police training manuals advise looking at verbal and nonverbal clues (Kassin, 2013) [25]. One example is that open-minded suspects are more likely to provide responses that are brief, clear, and exact, but suspects with a guilty conscience are less likely to maintain eye contact with the interrogator. While it's conceivable that these broad guidelines have some value, it's also evident from the available data that not even the most trained interrogator can reliably uncover dishonesty based on the verbal and nonverbal clues given by suspects (Kelly, Christopher & Miller, Jeaneé & Redlich) [29]. Furthermore, Vrij (2000) [30] asserts that the majority of dishonest individuals or liars will ultimately be detected since it becomes harder and harder to lie, and others still haven't Figure out how

to avoid being caught.

1.10 PROBLEM STATEMENT

Despite significant advances in affective computing, existing emotion recognition systems in educational environments predominantly rely on unimodal data and lack robustness under real-world conditions. The absence of effective multimodal fusion strategies and adaptive optimization mechanisms limits their accuracy and generalizability. Furthermore, current models fail to adequately capture the complex interplay between visual, auditory, and physiological emotional cues. This research addresses these limitations by proposing an optimized multimodal emotion recognition framework using deep learning and hybrid BAT–ROA optimization techniques.

1.11 MOTIVATIONS

The impetus for using multimodal emotion detection via deep learning networks and fusion approaches is to develop systems that are more precise, resilient, and contextually cognizant in interpreting and reacting to human emotions. Human emotions are expressed via facial expressions, voice tones, gestures, body movements, and posture. An automated system proficient in recognizing these emotions might transform several fields, such as video games, human-computer interaction, robotics, instructional software, animations, car safety, and affective computing. Consequently, creating a dependable, real-time emotion identification system is an essential objective, with its many applications necessitating comprehensive investigation. A hopeful consequence is the development of increasingly advanced robots capable of comprehending human emotions, resulting in substantial progress in this domain.

- **Richer Emotion Understanding:** Emotions are intricate and often conveyed via several modalities, including facial expressions, verbal communication, and written material. Integrating information from many modalities enables a more thorough and nuanced comprehension of an individual's emotional state.

- **Compensating for Ambiguities:** Every channel has inherent limits and may sometimes fail to effectively communicate emotions. Facial expressions may be confusing owing to cultural disparities, and written material may lack contextual

indicators. The integration of modalities mitigates these limitations, enhancing the overall dependability of emotion recognition.

- **Contextual Relevance:** Diverse modalities may provide context that improves the precision of emotion recognition. By integrating vocal intonations and facial expressions, the system may more effectively differentiate between sarcasm and sincerity in an individual's conversation.

- **Improved Robustness:** Unimodal emotion identification may be vulnerable to noise and unpredictability. By integrating various senses, the system enhances its resilience to fluctuations induced by variables like as lighting conditions, background noise, and individual disparities.

- **Personalization and Adaptation:** Multimodal methodologies enable systems to tailor and customize their replies according on the user's emotional condition. This fosters more empathic interactions, enhancing user engagement and pleasure.

1.12 CONTRIBUTIONS OF THESIS

This thesis offers a collection of novel contributions in the field of emotion recognition. This study presents a novel deep-learning method based on the BROA to facilitate face emotion identification using several data modalities. The BROA optimization method integrates the Bat method (BA) with the Rider Optimization Algorithm (ROA). This integration seeks to improve the precision of emotion identification by adeptly using facial pictures, EEG data, and physiological signals as components of the multimodal input.

The suggested approach aims to enhance the accuracy of multimodal emotion identification. Moreover, the study examines the ideal arrangement of deep learning networks. This study seeks to effectively identify essential characteristics that communicate human emotions via three modalities: signals, facial expressions, and body movements. The primary goal is to accurately categorize emotions, whether examining different modalities alone or in together.

An effective approach to emotion identification is achieved by using the suggested BROA alongside a multimodal signal. The recognition process includes several modalities, such as face pictures, EEG data, and physiological signals. Facial picture characteristics are

processed using a DBN, while data obtained from EEG and physiological signals are processed using a Deep RNN classifier.

This is a novel method for feature-level fusion, using multimodal compact bilinear pooling to integrate emotional signals from diverse sources. It illustrates its effectiveness in two particular contexts: the amalgamation of face and signal aspects and the integration of facial and bodily movement features. Empirical evidence from studies confirms the system's exceptional capacity to accurately recognize emotions from spatio-temporal data. Moreover, the method demonstrates its proficiency in efficiently accommodating various types of multimodal streams.

1.13 OBJECTIVE

1. To construct a multimodal method for recognizing emotions that incorporates facial expression analysis, physiological signals, and contextual data.

2. To apply ANN and CNN models for accurate and real-time emotion detection.

3. To explore the application of forensic psychology principles in interpreting emotional cues and detecting conflicts.

4. To evaluate the efficiency of the system in practical educational environments.

5. To Experimental and computational studies are used to recognize people's emotions by considering multiple modalities.

.

1.14 SCOPE OF THE STUDY

The scope of this research is confined to the development and evaluation of a multimodal emotion recognition system tailored for educational settings. The study focuses on integrating facial expressions, speech signals, and textual cues using machine learning and deep learning techniques. Psychological interpretations are supported through forensic psychology concepts to improve emotional understanding. Experimental validation is conducted using publicly available datasets and simulated classroom environments. The scope does not extend to clinical diagnosis or real-time deployment across large-scale educational institutions.

1.15 LIMITATIONS OF THE STUDY

Despite its contributions, the present study has certain limitations:

1. The analysis is limited to selected datasets, which may not fully represent real classroom diversity.
2. Real-time deployment constraints such as hardware costs and latency are not addressed.
3. Emotional states are inferred computationally and may not capture deep psychological nuances.
4. Cultural and linguistic variations among learners are only partially considered.

1.16 RESEARCH GAP

A critical review of existing literature (2020–2024) reveals that most emotion recognition studies in educational contexts emphasize either unimodal or basic multimodal approaches without optimized feature selection or psychological grounding. Limited research integrates forensic psychology with AI-driven emotion models to enhance interpretability and reliability. Additionally, existing frameworks lack robust optimization techniques to handle noisy educational data and contextual variability. This study addresses these gaps by proposing an optimized multimodal AI-based framework reinforced with forensic psychological perspectives for accurate emotion recognition in educational environments.

CHAPTER – 2
LITERATURE REVIEW

2.1 INTRODUCTION

Emotion recognition has emerged as a significant research domain within the broader field of affective computing, aiming to enable machines to interpret, understand, and respond to human emotional states. Human emotions are complex, multidimensional phenomena that are expressed through a variety of observable and latent modalities, including facial expressions, speech signals, physiological responses, and behavioral patterns. The increasing integration of artificial intelligence (AI) into domains such as education, healthcare, security, and human–computer interaction has intensified the need for accurate and robust emotion recognition systems.

Traditionally, emotion recognition approaches have relied on unimodal techniques, where a single source of emotional information—such as facial expressions or speech—is analyzed independently. Early research primarily focused on handcrafted feature extraction methods combined with classical machine learning algorithms. For instance, several studies have explored the use of Mel-Frequency Cepstral Coefficients (MFCCs) for speech-based emotion recognition, where features are extracted from audio signals and classified using models such as Support Vector Machines (SVMs) or neural networks [31], [32]. Similarly, facial expression recognition techniques have utilized geometric and appearance-based features, often enhanced through preprocessing methods such as filtering and transformation, followed by classification using machine learning or deep learning frameworks [33], [34].

Despite the advancements in unimodal approaches, these methods face significant limitations due to their inability to capture the full spectrum of human emotional expression. Emotions are inherently multimodal, and relying on a single modality often leads to ambiguity and reduced accuracy, especially in real-world scenarios where noise, occlusion, and variability are prevalent. For example, facial expressions alone may not fully represent emotional states under varying lighting conditions or occlusions, while speech signals may be affected by background noise or speaker variability. This has motivated researchers to explore additional modalities, including physiological signals such as electroencephalography (EEG), which provide insights into internal emotional

states by analyzing brain activity [35], [38].

The use of physiological signals has introduced a new dimension to emotion recognition research. EEG-based approaches leverage the brain's electrical activity to detect emotional responses with high temporal resolution. Studies have demonstrated the effectiveness of combining spatial and temporal feature extraction techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to analyze EEG signals and improve recognition accuracy [35], [48]. Furthermore, the development of publicly available datasets such as DEAP, DREAMER, and SEED has facilitated standardized evaluation and comparison of emotion recognition methods, enabling researchers to validate their models across diverse experimental conditions [38], [46], [47].

In addition to physiological signals, behavioral cues such as body gestures and skeletal movements have also been investigated for emotion recognition. Techniques utilizing motion capture systems and depth sensors, such as Kinect, have demonstrated the potential of gesture-based emotion recognition by analyzing joint coordinates and movement patterns [40]. These approaches highlight the importance of incorporating dynamic and contextual information into emotion recognition systems.

Recent advancements in deep learning have significantly transformed the landscape of emotion recognition. Deep neural networks, including CNNs, RNNs, and their variants, have enabled automatic feature extraction and hierarchical representation learning from raw data. These models have been successfully applied across different modalities, including facial images, speech signals, and physiological data, leading to improved performance compared to traditional methods. For example, deep CNN architectures have been employed for facial emotion recognition, while LSTM and GRU networks have been used to capture temporal dependencies in speech and EEG signals [33], [48], [50].

However, even with the adoption of deep learning techniques, unimodal systems continue to exhibit limitations in handling real-world variability and complexity. This has led to the emergence of multimodal emotion recognition (MER), which integrates multiple sources of emotional information to achieve more accurate and robust predictions. Multimodal approaches combine visual, auditory, and physiological signals to exploit complementary information across modalities, thereby addressing the shortcomings of individual modalities. Studies have shown that multimodal systems generally outperform unimodal systems in terms of recognition accuracy and reliability [67], [75].

The integration of multimodal data introduces new challenges, particularly in terms of data

fusion, synchronization, and feature representation. Various fusion strategies, including early fusion, late fusion, and hybrid approaches, have been proposed to effectively combine information from different modalities. Each approach presents its own advantages and limitations, depending on the application context and data characteristics. Additionally, the presence of noise, missing modalities, and domain variability further complicates the design of robust multimodal systems [69], [74].

Another critical aspect of emotion recognition research is the optimization of model performance. The high dimensionality of feature spaces and the complexity of deep learning architectures necessitate the use of advanced optimization techniques. Metaheuristic algorithms, such as Particle Swarm Optimization (PSO), Grey Wolf Optimization (GWO), and other evolutionary approaches, have been employed to optimize feature selection, model parameters, and fusion strategies [49], [50]. These techniques help improve classification accuracy, convergence speed, and generalization capability.

In recent years, there has been a growing interest in applying emotion recognition systems in real-world applications, particularly in educational environments. Emotion-aware systems can enhance learning experiences by adapting to students' emotional states, thereby improving engagement, motivation, and learning outcomes [66]. Similarly, applications in healthcare, security, and human–computer interaction have demonstrated the practical significance of emotion recognition technologies.

Despite the substantial progress in this field, several challenges remain. Issues such as data scarcity, cross-cultural variability, ethical concerns, privacy preservation, and model generalization continue to hinder the widespread deployment of emotion recognition systems. Furthermore, the lack of standardized frameworks for multimodal integration and the limited use of advanced optimization techniques highlight the need for further research.

In summary, emotion recognition has evolved from traditional unimodal approaches to advanced multimodal systems driven by deep learning and optimization techniques. While significant advancements have been achieved, the complexity of human emotions and the challenges associated with real-world deployment necessitate continued research and innovation in this domain.

Recent advancements in affective computing have been significantly influenced by the rapid development of deep learning and multimodal learning paradigms. Contemporary studies emphasize the role of large-scale multimodal datasets and advanced architectures in improving emotion recognition performance across diverse application domains. In

particular, recent surveys and systematic reviews highlight the transition from traditional machine learning approaches to deep learning-driven multimodal frameworks capable of handling complex emotional representations in real-world environments [67], [69]. Furthermore, the integration of artificial intelligence in educational and interactive systems has reinforced the importance of emotion-aware models for adaptive and intelligent decision-making [66].

2.2 FACIAL EMOTION RECOGNITION TECHNIQUES

Facial emotion recognition (FER) represents one of the most extensively researched and foundational areas within affective computing. Human facial expressions serve as one of the most direct and universal means of conveying emotions, enabling systems to interpret affective states through visual cues. Over the years, FER techniques have evolved significantly from traditional handcrafted feature-based methods to sophisticated deep learning-based architectures capable of learning complex spatial representations from large-scale datasets.

Early approaches to facial emotion recognition primarily relied on geometric and appearance-based feature extraction techniques. These methods focused on identifying key facial landmarks such as the eyes, eyebrows, nose, and mouth, and analyzing their relative positions and movements. For instance, approaches utilizing image enhancement techniques such as discrete wavelet transform (DWT) combined with fuzzy logic have been employed to improve facial feature extraction prior to classification [34]. Similarly, classical feature descriptors such as Gabor filters, Haar-like features, and ORB (Oriented FAST and Rotated BRIEF) have been widely used to capture texture and structural information from facial images [36]. These features are typically fed into machine learning classifiers such as Support Vector Machines (SVMs), decision trees, or ensemble models for emotion classification.

Despite their effectiveness in controlled environments, handcrafted feature-based approaches suffer from several limitations. They are highly sensitive to variations in illumination, pose, occlusion, and individual differences in facial expressions. Moreover, the manual design of features often fails to capture the complex and subtle variations associated with emotional expressions. These challenges have led to the adoption of deep learning techniques, particularly convolutional neural networks (CNNs), which enable

automatic feature learning directly from raw image data.

The introduction of CNN-based models has significantly improved the performance of facial emotion recognition systems. CNNs are capable of learning hierarchical representations of facial features, ranging from low-level edges and textures to high-level semantic patterns associated with specific emotions. For example, deep learning approaches have been employed to model the correlation between different emotional expressions using techniques such as correlation emotion label distribution learning, which assigns probabilistic weights to multiple emotions associated with a single expression [33]. This approach addresses the ambiguity inherent in facial expressions, where a single expression may correspond to multiple emotional states.

In addition to static image analysis, recent research has emphasized the importance of capturing temporal dynamics in facial expressions. Emotions are not static but evolve over time, and the ability to model these temporal variations is crucial for accurate emotion recognition. Techniques that combine spatial feature extraction with temporal modeling, such as CNNs integrated with recurrent neural networks (RNNs) or Long Short-Term Memory (LSTM) networks, have been proposed to capture both spatial and temporal information. These hybrid architectures enable the system to analyze sequences of facial frames, thereby improving recognition accuracy in dynamic scenarios.

Another important aspect of FER is the use of Facial Action Units (FAUs), which are based on the Facial Action Coding System (FACS). FAUs represent the activation of specific facial muscles that correspond to different expressions. By analyzing combinations of action units, systems can achieve a more interpretable and physiologically grounded understanding of emotions. Research has shown that incorporating both static and dynamic facial features, along with action unit analysis, can significantly enhance the robustness of emotion recognition systems [71]. Moreover, deep learning models have been used to automatically learn representations aligned with these action units, further improving classification performance.

The availability of benchmark facial expression datasets has played a crucial role in advancing FER research. Datasets such as CK+, Oulu-CASIA, KDEF, FER2013, and MMI provide labeled facial images and videos representing various emotional states under controlled and semi-controlled conditions [34], [36]. These datasets have enabled researchers to train and evaluate models systematically, facilitating comparative analysis of different approaches. However, many of these datasets are limited in terms of diversity,

as they often contain posed expressions captured in laboratory environments. This limitation affects the generalizability of models when applied to real-world scenarios.

To address these challenges, recent studies have focused on developing FER systems capable of operating in unconstrained or "in-the-wild" environments. These environments introduce additional complexities such as varying lighting conditions, background clutter, occlusions, and diverse facial orientations. Advanced deep learning architectures, including ResNet, VGGNet, and Inception-based models, have been employed to improve robustness under such conditions. Techniques such as data augmentation, normalization, and transfer learning have also been widely used to enhance model generalization.

Another emerging trend in FER research is the integration of edge computing and real-time processing capabilities. With the increasing demand for real-time emotion recognition in applications such as intelligent tutoring systems, surveillance, and human–computer interaction, there is a need for lightweight and efficient models. Studies have explored the use of edge devices and embedded systems to perform facial emotion recognition with reduced computational complexity while maintaining acceptable accuracy levels [72]. These approaches enable deployment in resource-constrained environments such as mobile devices and IoT systems.

In addition to performance improvements, recent research has also highlighted the importance of interpretability and explainability in FER systems. Understanding how a model arrives at a particular prediction is crucial for building trust and ensuring reliability, especially in sensitive applications such as healthcare and education. Techniques such as attention mechanisms and visualization of feature maps have been used to identify the regions of the face that contribute most to emotion classification.

Furthermore, FER has found applications across a wide range of domains, including education, healthcare, security, and entertainment. In educational settings, emotion recognition systems can be used to monitor students' engagement and adapt learning content accordingly [66]. In healthcare, FER can assist in the diagnosis and monitoring of mental health conditions by analyzing patients' emotional responses. Similarly, in security and surveillance, FER can be used to detect suspicious behavior and assess crowd emotions.

Despite the significant progress achieved in facial emotion recognition, several challenges remain. One of the primary challenges is the inherent variability and subjectivity of human emotions, which makes it difficult to define clear boundaries between different emotional

states. Additionally, cultural differences in emotional expression can affect the performance of FER systems trained on specific datasets. Privacy concerns and ethical considerations also pose challenges for the widespread adoption of FER technologies, particularly in applications involving continuous monitoring and data collection.

Recent studies in facial emotion recognition have focused on improving robustness under unconstrained conditions using deep learning architectures. Comprehensive surveys highlight the effectiveness of convolutional neural networks and hybrid deep learning models in capturing complex facial features and improving classification accuracy [72]. Additionally, recent research emphasizes the role of real-world datasets and domain adaptation techniques in enhancing generalization across diverse environments. These approaches address challenges such as illumination variation, occlusion, and pose differences, which are critical for practical deployment of FER systems [139].

In summary, facial emotion recognition has evolved from traditional feature-based approaches to advanced deep learning models capable of handling complex and dynamic facial expressions. While substantial improvements have been achieved in terms of accuracy and robustness, challenges related to generalization, interpretability, and real-world deployment continue to drive ongoing research in this field. The integration of FER with other modalities, such as speech and physiological signals, represents a promising direction for developing more comprehensive and reliable emotion recognition systems.

2.3 SPEECH / AUDIO EMOTION RECOGNITION TECHNIQUES

Speech or audio-based emotion recognition constitutes a critical component of affective computing, as vocal expressions provide rich information about an individual's emotional state. Unlike facial expressions, which may be consciously controlled or suppressed, speech signals often contain involuntary emotional cues embedded in tone, pitch, rhythm, and intensity. These characteristics make speech a highly valuable modality for emotion recognition, particularly in scenarios where visual information is unavailable or unreliable.

Early research in speech emotion recognition primarily focused on the extraction of handcrafted acoustic features, followed by classification using traditional machine learning techniques. Among these features, Mel-Frequency Cepstral Coefficients (MFCCs) have been widely adopted due to their effectiveness in capturing the perceptual characteristics of human auditory processing. Studies have demonstrated that MFCC features, along with their dynamic variants such as delta and delta-delta coefficients, can effectively represent

emotional variations in speech signals [31]. These features are typically used in conjunction with classifiers such as Support Vector Machines (SVMs), Gaussian Mixture Models (GMMs), and Hidden Markov Models (HMMs) to perform emotion classification.

In addition to MFCCs, other acoustic features such as pitch, energy, spectral centroid, formants, and zero-crossing rate have been explored to enhance emotion recognition performance. These features capture various aspects of speech, including prosodic, spectral, and temporal characteristics. For example, variations in pitch and energy can indicate emotional states such as anger or excitement, while changes in speech rate and rhythm may reflect sadness or fatigue. The combination of multiple feature types often leads to improved recognition accuracy by providing a more comprehensive representation of the speech signal.

To address the limitations of traditional feature extraction methods, researchers have introduced advanced preprocessing and feature transformation techniques. For instance, discriminant analysis has been applied to MFCC features to reduce inter-class similarity and enhance class separability prior to classification [32]. Similarly, wavelet transforms and frequency-domain analysis techniques have been used to decompose speech signals into multiple sub-bands, enabling the extraction of more discriminative features. These approaches aim to improve the robustness of emotion recognition systems under varying acoustic conditions.

With the advent of deep learning, speech emotion recognition has undergone a significant transformation. Deep neural networks, including Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Gated Recurrent Units (GRUs), have been widely adopted for modeling temporal dependencies in speech signals. Unlike traditional methods, which rely on fixed-length feature vectors, these models can process sequential data and capture long-term dependencies, making them particularly suitable for analyzing speech dynamics.

RNN-based models have demonstrated superior performance in capturing temporal patterns associated with emotional speech. For example, LSTM networks are capable of retaining information over long sequences, allowing them to model variations in pitch, tone, and intensity over time. Similarly, GRU-based models provide a computationally

efficient alternative to LSTMs while maintaining comparable performance. These models are often combined with feature extraction techniques such as MFCCs to create hybrid architectures that leverage both handcrafted and learned features.

In addition to RNN-based approaches, deep convolutional neural networks (CNNs) have also been applied to speech emotion recognition by treating spectrograms as image-like representations of audio signals. Spectrograms, which visualize the frequency content of a signal over time, enable CNNs to learn spatial patterns associated with different emotional states. This approach bridges the gap between image-based and audio-based analysis, allowing researchers to leverage advances in computer vision for speech processing tasks.

Another significant advancement in speech emotion recognition is the use of hybrid deep learning architectures, which combine multiple neural network models to capture different aspects of speech signals. For example, CNN-RNN hybrid models integrate spatial feature extraction with temporal modeling, enabling the system to learn both local and sequential patterns. These architectures have shown improved performance compared to standalone models, particularly in complex and noisy environments.

Despite these advancements, speech emotion recognition faces several challenges, particularly in real-world applications. One of the primary challenges is the variability in speech signals caused by factors such as speaker differences, language, accent, and environmental noise. For instance, the same emotional state may be expressed differently by different individuals, making it difficult to develop generalized models. Additionally, background noise and recording conditions can significantly affect the quality of speech signals, leading to reduced recognition accuracy.

To mitigate these challenges, researchers have explored the use of noise reduction and signal enhancement techniques. Methods such as filtering, normalization, and data augmentation are commonly employed to improve the robustness of speech emotion recognition systems. Furthermore, the development of large-scale and diverse datasets has enabled the training of more generalized models capable of handling real-world variability.

Another important aspect of speech emotion recognition is the use of speaker-independent models, which aim to generalize across different individuals. Traditional models often exhibit high accuracy when trained and tested on the same speaker but perform poorly in

cross-speaker scenarios. To address this issue, techniques such as domain adaptation, transfer learning, and subject-independent feature extraction have been proposed [32]. These approaches aim to reduce the dependence on specific speaker characteristics and improve the generalization capability of emotion recognition systems.

In addition to standalone speech-based approaches, audio signals have also been integrated into multimodal emotion recognition systems, where they complement other modalities such as facial expressions and physiological signals. Speech provides valuable contextual information that may not be captured by visual or physiological data alone. For example, vocal tone and intonation can convey subtle emotional nuances that are not visible in facial expressions. The integration of speech with other modalities has been shown to significantly enhance overall system performance.

Recent research has also explored the application of speech emotion recognition in various domains, including healthcare, education, and human–computer interaction. In healthcare, speech analysis can be used to detect mental health conditions such as depression and anxiety by identifying changes in vocal patterns. In educational settings, speech-based emotion recognition can be used to assess student engagement and provide personalized feedback. Similarly, in human–computer interaction, emotion-aware systems can adapt their responses based on the user's emotional state, improving user experience and interaction quality.

Furthermore, advancements in deep learning optimization and model training techniques have contributed to improved performance in speech emotion recognition. Techniques such as learning rate scheduling, early stopping, and regularization are commonly used to enhance model convergence and prevent overfitting. Additionally, the use of large-scale pre-trained models and transfer learning has enabled researchers to leverage knowledge from related tasks, further improving recognition accuracy.

Despite these advancements, several challenges remain in the field of speech emotion recognition. Issues such as data imbalance, lack of standardized evaluation protocols, and limited availability of labeled datasets continue to hinder progress. Moreover, ethical concerns related to privacy and data security must be addressed to ensure the responsible use of emotion recognition technologies.

Recent advancements in speech emotion recognition have been driven by improvements in deep learning-based acoustic modeling and feature representation techniques. Studies demonstrate the effectiveness of neural structured learning and deep architectures in improving robustness to noise and variability in speech signals [32]. Additionally, recent reviews on speech-based emotion recognition highlight the importance of selecting discriminative acoustic features and robust classification strategies to improve performance in real-world scenarios [127]. These developments have contributed to more reliable and scalable speech emotion recognition systems.

In summary, speech emotion recognition has evolved from traditional feature-based methods to advanced deep learning architectures capable of modeling complex temporal dynamics in speech signals. While significant progress has been made, challenges related to variability, noise, and generalization continue to drive ongoing research. The integration of speech with other modalities in multimodal systems represents a promising direction for achieving more accurate and robust emotion recognition in real-world applications.

2.4 MULTIMODAL EMOTION RECOGNITION APPROACHES

Multimodal emotion recognition (MER) has emerged as a highly effective approach for accurately identifying human emotions by integrating multiple sources of information, such as facial expressions, speech signals, physiological responses, and behavioral cues. Unlike unimodal systems, which rely on a single modality and are often limited by noise, ambiguity, or incomplete information, multimodal approaches leverage the complementary nature of different modalities to provide a more comprehensive and robust understanding of emotional states.

Human emotions are inherently complex and are expressed through a combination of verbal and non-verbal signals. For instance, while facial expressions convey visible emotional cues, speech signals capture vocal nuances such as tone, pitch, and rhythm, and physiological signals reflect internal emotional responses. Relying on a single modality often leads to incomplete or ambiguous interpretations, particularly in real-world scenarios where environmental factors and individual variability play a significant role. This limitation has driven the development of multimodal frameworks that integrate diverse data sources to enhance emotion recognition performance.

Early research in multimodal emotion recognition focused on combining two or more

modalities using relatively simple fusion strategies. These approaches typically involved feature-level (early) fusion or decision-level (late) fusion. In feature-level fusion, features extracted from different modalities are concatenated into a single feature vector before classification. This approach enables the model to learn correlations between modalities but may suffer from high dimensionality and synchronization challenges. In contrast, decision-level fusion involves combining the outputs of individual classifiers trained on different modalities. While this approach is more flexible and easier to implement, it may fail to capture inter-modal relationships effectively.

Recent advancements in deep learning have led to the development of more sophisticated fusion strategies, including hybrid and intermediate fusion techniques. These approaches integrate modalities at multiple levels within the network, allowing for more effective interaction between features. For example, intermediate fusion methods combine features from different modalities at specific layers of deep neural networks, enabling the model to learn joint representations that capture both modality-specific and cross-modal information [67]. Such approaches have demonstrated improved performance compared to traditional fusion techniques.

The integration of multimodal data requires careful consideration of data synchronization and alignment. Different modalities often operate at varying temporal resolutions; for instance, video data may be captured at frame rates of 25–30 frames per second, while audio signals are sampled at much higher frequencies. Aligning these modalities is crucial for ensuring that corresponding features represent the same temporal segment of emotional expression. Techniques such as temporal segmentation, windowing, and interpolation are commonly used to achieve synchronization across modalities.

In addition to synchronization, multimodal systems must address the challenge of heterogeneous data representation. Each modality has distinct characteristics and requires specialized feature extraction techniques. For example, CNNs are commonly used for extracting spatial features from facial images, while RNNs or LSTM networks are employed for modeling temporal dependencies in speech and physiological signals. The integration of these diverse feature types necessitates the use of architectures capable of handling heterogeneous inputs and learning unified representations.

Several studies have highlighted the effectiveness of multimodal emotion recognition in improving classification accuracy. By combining visual, auditory, and physiological cues, these systems can overcome the limitations of individual modalities. For instance, facial expressions may be ambiguous or suppressed in certain situations, but speech signals can provide additional context to disambiguate emotional states. Similarly, physiological signals such as EEG can capture internal emotional responses that are not externally visible [69], [75].

The development of multimodal datasets has played a crucial role in advancing MER research. Datasets such as DEAP, DREAMER, and MPED include multiple modalities, including EEG, ECG, GSR, audio, and video, enabling researchers to evaluate multimodal systems under controlled conditions [43], [46]. These datasets provide synchronized recordings of different modalities, facilitating the development and benchmarking of multimodal models. However, many existing datasets are limited in size and diversity, which can affect the generalization of trained models.

Recent research has also explored the use of deep multimodal learning frameworks, which leverage the power of deep neural networks to learn complex relationships between modalities. These frameworks often incorporate attention mechanisms, which allow the model to focus on the most relevant features from each modality. Attention-based models dynamically assign weights to different modalities or features, enabling the system to prioritize more informative cues and improve recognition performance.

Another important advancement in multimodal emotion recognition is the use of representation learning techniques, such as autoencoders and generative models, to learn compact and discriminative feature representations. These techniques aim to reduce the dimensionality of multimodal data while preserving essential information. For example, multimodal autoencoders can learn joint representations by encoding data from multiple modalities into a shared latent space, facilitating more effective fusion and classification.

Despite the advantages of multimodal approaches, several challenges remain. One of the primary challenges is the presence of missing or noisy modalities. In real-world scenarios, it is not always possible to obtain all modalities simultaneously. For example, video data may be unavailable due to camera failure, or audio signals may be corrupted by noise. To

address this issue, researchers have developed techniques for handling incomplete data, such as modality dropout, imputation, and robust fusion strategies that can operate with partial inputs.

Another challenge is the computational complexity associated with multimodal systems. The integration of multiple modalities increases the dimensionality of the data and the complexity of the model, leading to higher computational and memory requirements. This can limit the applicability of multimodal systems in real-time or resource-constrained environments. To mitigate this issue, researchers have explored the use of lightweight models, model compression techniques, and edge computing solutions to enable efficient deployment.

The issue of context awareness also plays a significant role in multimodal emotion recognition. Emotions are influenced by contextual factors such as environment, social interactions, and cultural background. Many existing models do not adequately account for these factors, leading to reduced performance in real-world applications. Incorporating contextual information into multimodal systems remains an open research challenge.

In addition to technical challenges, ethical and privacy concerns are increasingly important in the development of multimodal emotion recognition systems. The collection and analysis of sensitive data, such as facial images and physiological signals, raise concerns about user privacy and data security. Ensuring transparency, consent, and responsible use of data is essential for the widespread adoption of these technologies.

Multimodal emotion recognition has found applications in a wide range of domains. In education, emotion-aware systems can monitor student engagement and adapt instructional strategies to improve learning outcomes [66], [68]. In healthcare, multimodal systems can assist in diagnosing and monitoring mental health conditions by analyzing behavioral and physiological signals. In human–computer interaction, emotion recognition enables the development of adaptive interfaces that respond to user emotions, enhancing user experience and interaction quality.

Furthermore, recent studies have explored the use of contactless multimodal emotion recognition systems, which utilize non-invasive sensors to capture emotional cues without physical interaction. These systems are particularly useful in applications where user

comfort and convenience are important, such as remote learning and telemedicine [75]. However, achieving high accuracy in contactless systems remains challenging due to noise and variability in the captured data.

Recent research in multimodal emotion recognition has focused on integrating visual, vocal, and physiological signals to improve system robustness and accuracy. Comprehensive surveys demonstrate that multimodal approaches significantly outperform unimodal systems by leveraging complementary information across modalities [67], [69]. Furthermore, recent studies have explored advanced multimodal frameworks that incorporate facial expressions, speech, and body gestures to achieve more reliable emotion recognition in real-world environments [74]. Emerging research also highlights the importance of contactless multimodal systems for practical applications, emphasizing scalability and user convenience [75].

In summary, multimodal emotion recognition represents a significant advancement over unimodal approaches by integrating multiple sources of emotional information to achieve higher accuracy and robustness. The development of advanced fusion strategies, deep learning architectures, and multimodal datasets has contributed to substantial progress in this field. However, challenges related to data heterogeneity, synchronization, computational complexity, and real-world deployment continue to drive ongoing research. Addressing these challenges is essential for the development of reliable and scalable multimodal emotion recognition systems capable of operating effectively in diverse application domains.

2.5 OPTIMIZATION TECHNIQUES IN EMOTION RECOGNITION

Optimization plays a crucial role in enhancing the performance of emotion recognition systems, particularly in the context of deep learning and multimodal frameworks. As emotion recognition models become increasingly complex, involving high-dimensional feature spaces and multiple modalities, the need for efficient optimization techniques becomes more pronounced. Optimization techniques are employed to improve various aspects of model performance, including feature selection, parameter tuning, convergence speed, and generalization capability.

Traditional machine learning approaches relied on deterministic optimization methods, such as gradient descent and its variants, to minimize loss functions and update model

parameters. While these methods are effective for convex optimization problems, they often struggle with non-convex and high-dimensional search spaces commonly encountered in deep learning models. Issues such as local minima, slow convergence, and sensitivity to initial conditions can significantly affect model performance. To address these challenges, researchers have explored the use of metaheuristic optimization algorithms, which are inspired by natural processes and provide robust solutions for complex optimization problems.

Among the widely used metaheuristic techniques, Particle Swarm Optimization (PSO), Genetic Algorithms (GA), and Grey Wolf Optimization (GWO) have been extensively applied in emotion recognition tasks. These algorithms operate by maintaining a population of candidate solutions and iteratively updating them based on specific rules that mimic natural behaviors. For instance, PSO simulates the social behavior of birds flocking or fish schooling, where particles adjust their positions based on their own experience and that of neighboring particles. Similarly, GA employs evolutionary principles such as selection, crossover, and mutation to evolve solutions over successive generations. These techniques have been used for tasks such as feature selection, hyperparameter tuning, and model optimization, leading to improved classification accuracy and reduced computational complexity [49], [50].

In the context of emotion recognition, feature selection is a critical optimization problem due to the high dimensionality of data, particularly in multimodal systems. Redundant or irrelevant features can degrade model performance and increase computational cost. Metaheuristic algorithms have been effectively used to identify optimal subsets of features that maximize classification accuracy while minimizing redundancy. For example, PSO-based feature selection methods have been applied to speech emotion recognition to select the most informative acoustic features, resulting in improved performance compared to traditional approaches.

Another important application of optimization techniques is in the tuning of hyperparameters of deep learning models. Parameters such as learning rate, number of hidden neurons, batch size, and network architecture significantly influence model performance. Manual tuning of these parameters is time-consuming and often suboptimal. Metaheuristic algorithms provide an automated approach to hyperparameter optimization

by exploring the search space efficiently and identifying optimal configurations. This leads to improved convergence behavior and better generalization of the model.

With the increasing adoption of multimodal emotion recognition systems, optimization techniques have also been applied to fusion strategies. Determining the optimal way to combine features or decisions from different modalities is a complex problem that can significantly impact system performance. Metaheuristic algorithms have been used to optimize fusion weights and strategies, enabling the system to dynamically adjust the contribution of each modality based on its relevance and reliability. This is particularly important in scenarios where certain modalities may be noisy or less informative.

In recent years, researchers have explored the use of hybrid optimization techniques, which combine multiple metaheuristic algorithms to leverage their complementary strengths. Hybrid approaches aim to balance exploration and exploitation, two fundamental aspects of optimization. Exploration refers to the ability to search broadly across the solution space, while exploitation focuses on refining solutions in promising regions. Achieving an optimal balance between these two aspects is essential for avoiding local optima and ensuring efficient convergence.

One such hybrid approach involves the combination of the Bat Algorithm (BA) and the Rider Optimization Algorithm (ROA). The Bat Algorithm is inspired by the echolocation behavior of bats and is known for its strong exploration capability. It uses parameters such as frequency, loudness, and pulse emission rate to navigate the search space and identify promising regions. On the other hand, the Rider Optimization Algorithm is inspired by the behavior of riders in a race, where different strategies are employed to reach the target efficiently. ROA is particularly effective in exploitation, as it refines solutions by moving them toward optimal regions using adaptive strategies.

The integration of BA and ROA into a hybrid optimization framework enables the system to benefit from both global exploration and local exploitation. In the context of emotion recognition, this hybrid approach can be used to optimize various components of the model, including feature selection, fusion weights, and network parameters. By combining the strengths of both algorithms, the hybrid BAT–ROA approach enhances convergence speed, improves classification accuracy, and reduces the likelihood of getting trapped in local

minima.

In addition to metaheuristic optimization, recent research has also explored the use of deep learning optimization techniques, such as adaptive learning rate methods (e.g., Adam, RMSProp) and regularization techniques (e.g., dropout, batch normalization). These methods improve the training process of neural networks by stabilizing gradients and preventing overfitting. While these techniques are effective, they are often used in conjunction with metaheuristic algorithms to achieve optimal performance.

Another emerging trend in optimization is the use of multi-objective optimization, where multiple performance metrics are optimized simultaneously. In emotion recognition, this may involve balancing accuracy, computational efficiency, and robustness. Multi-objective optimization techniques enable the development of models that perform well across multiple criteria, making them more suitable for real-world applications.

Despite the advantages of optimization techniques, several challenges remain. Metaheuristic algorithms can be computationally intensive, particularly when applied to large datasets and complex models. Additionally, the performance of these algorithms depends on the choice of parameters and initialization, which may require careful tuning. Hybrid optimization approaches, while effective, can further increase computational complexity.

Furthermore, the integration of optimization techniques into deep learning frameworks requires careful design to ensure compatibility and efficiency. The dynamic nature of deep learning models, combined with the stochastic behavior of metaheuristic algorithms, can introduce additional challenges in terms of stability and reproducibility.

In practical applications, optimization techniques have been successfully applied in various domains, including education, healthcare, and intelligent systems. For example, in intelligent tutoring systems, optimized emotion recognition models can adapt to students' emotional states in real time, enhancing learning outcomes. In healthcare, optimized models can improve the accuracy of emotion detection for mental health assessment and monitoring.

Recent studies emphasize the importance of optimization techniques in improving the

performance of deep learning-based emotion recognition systems. Advanced approaches incorporate evolutionary and hybrid optimization strategies to enhance feature selection, parameter tuning, and model convergence [49], [50]. Additionally, recent research highlights the growing role of intelligent optimization frameworks in improving computational efficiency and scalability, particularly in complex multimodal systems. These developments support the need for hybrid optimization techniques that can effectively balance exploration and exploitation in high-dimensional search spaces.

In summary, optimization techniques play a vital role in advancing emotion recognition systems by improving model performance, efficiency, and robustness. The use of metaheuristic algorithms, particularly hybrid approaches such as BAT–ROA, provides a powerful framework for addressing the challenges associated with high-dimensional and non-convex optimization problems. While significant progress has been made, ongoing research is required to address challenges related to computational complexity, parameter tuning, and real-world deployment. The integration of advanced optimization techniques with multimodal emotion recognition systems represents a promising direction for future research.

Table 2.1: Comparison of Emotion Recognition Datasets Used in Existing Studies

Ref. No.	Methodology / Model	Modality	Key Contribution	Limitations
[31]	SVM-based emotion detection	Audio	Uses MFCC features for speech emotion recognition	Limited accuracy; lacks temporal modeling
[32]	Neural structured learning	Audio	Improves robustness using deep learning	Computational complexity; unimodal limitation
[33]	Gaussian label distribution learning	Facial	Handles ambiguity in facial expressions	Sensitive to lighting conditions
[34]	Eyemap–Mouthmap + TensorFlow	Facial	Enhances facial feature extraction	Limited real-world adaptability
[35]	Attention-based EEG model	Physiological	Captures spatial-temporal brain signals	Requires specialized hardware

[48]	RNN-based EEG classification	Physiological	Models temporal dependencies effectively	High computational cost
[49]	Deep learning + feature statistics	Physiological	Improved feature extraction for emotion detection	No multimodal integration
[50]	BLSTM-LSTM + evolutionary algorithm	Physiological	Combines deep learning with optimization	Limited to EEG domain
[66]	AI in learning environments (Review)	Multimodal	Shows role of emotion recognition in education	Lacks implementation framework
[67]	Deep multimodal survey	Multimodal	Comprehensive review of multimodal techniques	No optimization strategies
[69]	Multimodal review (visual + vocal + physiological)	Multimodal	Highlights importance of multimodal fusion	Fusion complexity not resolved
[72]	Deep FER survey	Facial	Identifies challenges in real-world FER	Limited multimodal integration
[74]	Multimodal (face + speech + gesture)	Multimodal	Improves recognition accuracy using multiple modalities	High computational complexity
[75]	Contactless multimodal systems	Multimodal	Enables non-invasive emotion recognition	Noise sensitivity; scalability issues
[139]	Deep learning emotion visualization	Facial	Improves representation learning	Limited generalization
[140]	Multimodal deep learning survey	Multimodal	Covers speech, text, and facial integration	Lacks optimization focus

The comparative analysis reveals that while significant advancements have been made in emotion recognition, most existing studies are limited by their reliance on unimodal approaches or simplistic multimodal fusion techniques. Although deep learning models have improved feature extraction and classification performance, they often lack efficient optimization mechanisms. Furthermore, many multimodal systems suffer from high computational complexity and inadequate handling of heterogeneous data. The absence of hybrid optimization strategies and scalable architectures highlights a critical research gap,

which motivates the development of an optimized multimodal framework integrating advanced feature extraction with hybrid BAT–ROA optimization.

2.6 SUMMARY OF RESEARCH GAPS

Despite the extensive research and significant advancements in the field of emotion recognition, several critical gaps remain that limit the effectiveness, scalability, and real-world applicability of existing systems. These gaps arise from limitations in unimodal approaches, challenges in multimodal integration, insufficient optimization strategies, and issues related to generalization, interpretability, and deployment. Identifying and addressing these gaps is essential for the development of robust and efficient emotion recognition frameworks.

One of the primary research gaps lies in the dominance of unimodal emotion recognition approaches in earlier studies. A large portion of the literature focuses on analyzing a single modality, such as facial expressions or speech signals, independently. While these approaches have demonstrated reasonable performance under controlled conditions, they fail to capture the multifaceted nature of human emotions. Unimodal systems are particularly vulnerable to noise, occlusion, and environmental variability. For instance, facial emotion recognition systems may struggle with variations in lighting, pose, and facial occlusions, while speech-based systems are affected by background noise and speaker variability. This limitation highlights the need for more comprehensive approaches that integrate multiple modalities to improve robustness and accuracy.

Another significant gap is related to the inefficient integration of multimodal data. Although multimodal emotion recognition has gained attention, many existing approaches rely on simplistic fusion techniques that do not fully exploit the relationships between different modalities. Feature-level fusion methods often suffer from high dimensionality and synchronization issues, while decision-level fusion approaches fail to capture inter-modal dependencies effectively. Furthermore, many studies do not adequately address the challenges of temporal alignment and heterogeneous data representation, which are critical for accurate multimodal analysis. This indicates a need for more advanced fusion strategies that can effectively combine information from multiple modalities while preserving their unique characteristics.

The lack of robust optimization techniques in many emotion recognition systems represents

another important research gap. While deep learning models have improved feature extraction and classification capabilities, their performance heavily depends on the selection of optimal parameters and feature subsets. Many existing studies rely on standard optimization methods or manual tuning, which may not yield optimal results. The limited use of advanced metaheuristic algorithms for optimizing model parameters, feature selection, and fusion strategies restricts the potential performance of these systems. This gap underscores the importance of incorporating efficient optimization techniques, such as hybrid metaheuristic algorithms, to enhance model performance and convergence.

A further gap exists in the limited exploration of hybrid optimization approaches. Although individual metaheuristic algorithms such as Particle Swarm Optimization (PSO) and Genetic Algorithms (GA) have been applied in emotion recognition, the potential of hybrid approaches that combine multiple algorithms has not been fully explored. Hybrid optimization techniques can leverage the strengths of different algorithms to achieve a better balance between exploration and exploitation, leading to improved performance. The absence of such approaches in existing literature highlights an opportunity for developing more effective optimization frameworks.

Another critical challenge is the lack of generalization and robustness in real-world environments. Many emotion recognition models are trained and evaluated on controlled datasets, which do not accurately represent real-world conditions. As a result, these models often fail to perform well when deployed in practical scenarios involving diverse populations, varying environmental conditions, and unpredictable inputs. The limited availability of large-scale, diverse, and annotated datasets further exacerbates this issue. Addressing this gap requires the development of models that can generalize across different domains and adapt to varying conditions.

The issue of data scarcity and imbalance also presents a significant research gap. Emotion recognition datasets often suffer from imbalanced class distributions, where certain emotions are underrepresented. This imbalance can lead to biased models that perform poorly on minority classes. Additionally, the collection and annotation of emotion data are time-consuming and resource-intensive processes, resulting in limited availability of high-quality datasets. Techniques such as data augmentation, transfer learning, and synthetic data generation have been explored to address these challenges, but further research is

needed to develop more effective solutions.

Another important gap is the lack of interpretability and explainability in emotion recognition models. Deep learning models, while highly effective, often function as “black boxes,” making it difficult to understand how decisions are made. This lack of transparency can hinder trust and adoption, particularly in sensitive applications such as healthcare and education. There is a need for methods that can provide insights into model behavior and highlight the features or modalities that contribute to specific predictions.

The computational complexity and scalability of emotion recognition systems also pose challenges for real-time deployment. Multimodal systems, in particular, require significant computational resources due to the integration of multiple data streams and complex model architectures. This limits their applicability in resource-constrained environments such as mobile devices and edge computing platforms. Developing lightweight and efficient models that can operate in real time without compromising accuracy remains an open research problem.

In addition to technical challenges, ethical and privacy concerns represent a significant gap in the field of emotion recognition. The collection and analysis of sensitive data, including facial images, speech signals, and physiological measurements, raise concerns about user privacy and data security. Ensuring compliance with ethical standards and developing privacy-preserving techniques are essential for the responsible deployment of emotion recognition systems.

Furthermore, there is a lack of standardized evaluation frameworks and benchmarking protocols in emotion recognition research. Different studies use varying datasets, evaluation metrics, and experimental setups, making it difficult to compare results and assess the effectiveness of different approaches. Establishing standardized benchmarks and evaluation criteria would facilitate more consistent and meaningful comparisons across studies.

Another gap is the limited application-specific customization of emotion recognition systems. While many studies focus on general-purpose emotion recognition, fewer efforts have been made to tailor these systems for specific applications such as education, healthcare, or human–computer interaction. Application-specific requirements, such as

real-time performance, interpretability, and domain-specific features, need to be considered to ensure practical usability.

Recent literature indicates that, despite the adoption of deep learning and multimodal approaches, several challenges remain unresolved. Studies highlight issues related to generalization, real-time implementation, and computational complexity in emotion recognition systems [69], [75]. Furthermore, while recent multimodal frameworks have improved accuracy, they often lack efficient optimization mechanisms and struggle with scalability in practical applications. These limitations reinforce the need for advanced hybrid optimization-based multimodal frameworks capable of addressing both performance and efficiency challenges.

In summary, the literature reveals several key research gaps, including the limitations of unimodal approaches, challenges in multimodal integration, insufficient optimization techniques, lack of hybrid optimization strategies, and issues related to generalization, data availability, interpretability, and scalability. Addressing these gaps is essential for advancing the field of emotion recognition and developing systems that are both accurate and practical for real-world applications. The proposed research aims to address these challenges by developing a multimodal emotion recognition framework optimized using a hybrid Bat Algorithm and Rider Optimization Algorithm (BAT–ROA), thereby improving performance, robustness, and applicability across diverse scenarios.

CHAPTER – 3
DESIGN AND METHODS

3.1 OVERVIEW

In this section, we lay out the steps used to collect data, clean it up, and analyze it so that we could build a classification model based on the methods we selected. This thesis employs a controlled experimental design in which a number of variables are maintained throughout. Alternatively, the second set of variables is experimentally being monitored. Various dataset ratios are used in various investigations. We conclude with the presentation and discussion of assessment methodologies used to determine the efficacy of the developed model.

Table 3.1: Hardware and Software Environment

Component	Specification
Processor	Intel Core i7 / AMD Ryzen 7
RAM	16 GB
Storage	512 GB SSD
GPU	NVIDIA GTX 1660 / RTX 3060
Operating System	Windows 10 / Ubuntu 20.04
Programming Language	Python 3.8
Deep Learning Framework	TensorFlow, Keras
Machine Learning Libraries	Scikit-learn, NumPy, Pandas
Visualization Tools	Matplotlib, Seaborn

Table 3.2: Dataset Distribution

Dataset Split	Number of Samples	Percentage
Training Set	70%	70%
Validation Set	15%	15%
Testing Set	15%	15%
Total	100%	100%

3.2 PROPOSED SYSTEM ARCHITECTURE

This thesis employs an experimental research methodology. This section delineates the methodologies and strategies used in this study to attain the thesis's purpose. The four fundamental processes of data pre-processing for emotion identification are face recognition, feature extraction, and classification. Figure 3.1 illustrates the major tasks of the study and the subsequent process flow. This paper proposes a feature extraction and classification technique for face emotion recognition using the VGG16, Google Net, and ResNet50 models. This study assigned seven distinct emotions—joy, grief, fury, surprise, fear, disdain, and neutrality—to photos from the dataset.

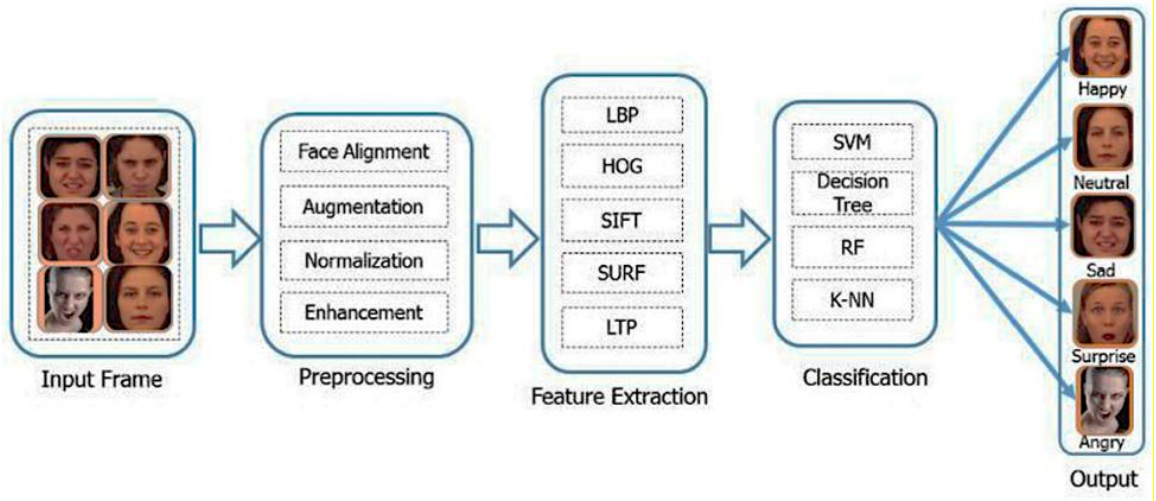


Figure 3.1: Flow of research

Research by Rozgić et al. presents a multimodal emotion identification approach using

facial, audio, and textual characteristics. The system architecture for this research is shown in Figure. Emotion recognition was performed at the sentence level, analyzing a total of 5,531 sentences. The used database was the University of Southern California's Interactive Emotional Motion Capture (USCIEMOCAP) corpus including audio and visual data. Facial characteristics were retrieved that denote vectors between various places on the face, grounded in Facial Animation Parameters (FAPs). Acoustic characteristics including F0, intensity, MFCCs, jitter, and shimmer were retrieved from each audio clip with audio analysis software such as Praat. Each sentence's text was acquired by ASR, followed by the extraction of two sets of textual characteristics: emotion and word stem features, utilizing the LIWC (Linguistic Inquiry and Word Count) and General Inquirer lexicons.

An ensemble of binary SVM classifiers was constructed after feature extraction. Each SVM classifier is trained on either the whole or a part of the available multimodal characteristics, as well as on a subset of potential categorized emotions, namely anger, pleasure, sadness, or neutrality. The proposed ensemble tree method surpasses two leading baseline classifiers across all categories of training features, with the most significant enhancement in classification accuracy observed for the ensemble tree utilizing all facial expression, speech, and text features, attaining 69.4% accuracy compared to 66.4% and 66.5% achieved by each respective baseline classifier.

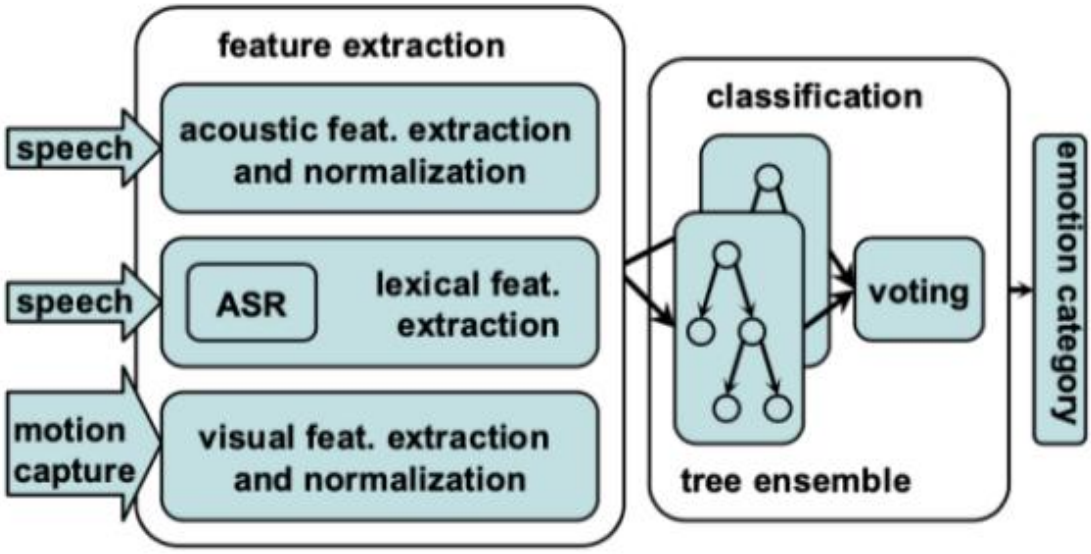


Figure 3.2: System architecture of multimodal emotion recognition using an ensemble tree of binary SVM classifiers

creates an emotion and sentiment identification system that uses many modalities to test the idea that merging them increases performance compared to single-modal systems. The six primary emotions identified by Ekman—anger, disgust, fear, pleasure, sorrow, and surprise—form the basis for the target emotions to be derived from each modality. Text characteristics like emotion (using the new lexicon EmoSentic Space) and sentiment polarity (using the SenticNet vocabulary) were extracted from the International Survey of Emotion Antecedents and Reactions (ISEAR) database. For the speech, we utilized the entry FACE database and extracted audio characteristics like MFCC and Spectral centroid using the JAudio audio analysis program. We utilized the CK+ database to store facial expressions, and we used the facial recognition program Luxand FSDK 1.74 to monitor facial characteristic points (FCPs) and extract features.

Figure 3.3 below shows the used FCPs for data facial feature extraction. The SVM classifier had the maximum classification accuracy after performing feature-level fusion, which included merging the features of all modalities into a single feature vector. This vector was then used to train multiple classifiers. This paper's suggested multimodal classifier outperformed state-of-the-art classifiers on two databases: the MMI database (72.10% accuracy) and the FABO database (61.21% accuracy), surpassing both databases' respective present state-of-the-art classifiers by 16.5% and 25.71%. In comparison to the face, vocal, and text unimodal classifiers, the multimodal classifier outperformed them all across all six emotion categories. Viewed in Figure 3.4 below, the authors have implemented their multimodal framework into a real-time multimodal emotion identification system. Another use of ASR in this system was to extract text from audio.

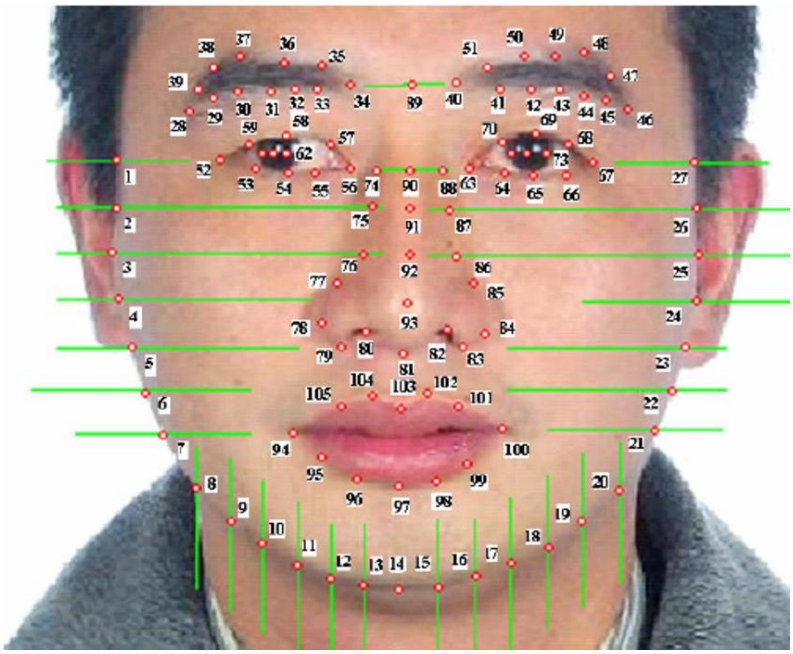


Figure 3.3: Facial Characteristic Points detected in a facial image

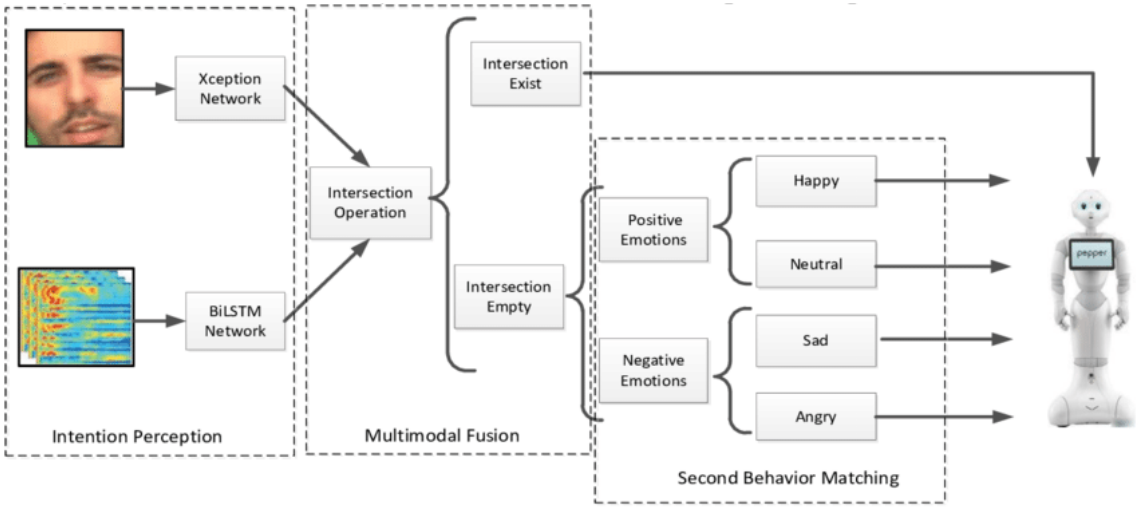


Figure 3.4: Real-time multimodal emotion recognition system architecture

As part of a cognitive robotics system, another research uses a multimodal approach to emotion detection. You can view the system's design in Figure 3.5 down below. As input, the system takes in audio and video data and uses it to extract features like facial expressions (e.g., brows, lips, and muscle movements) from video, voice quality and intensity from audio, and syntactic structure and meaning from transcripts of audio to create text features. In order to accomplish decision-level data fusion, the output of the classifiers used by each of these modalities is fed into the input layer of a Multilayer

Perceptron (MLP) neural network. Anger, contempt, fear, pleasure, neutral, sorrow, surprise, and disgust are the emotion classes for which the MLP produces predictions. With an average accuracy of 81.65% compared to the state-of-the-art model's 74.09%, our MLP model achieves superior classification accuracy. An additional 11.36% increase above the state-of-the-art—85.45%—in classification accuracy was achieved when this research experimented with adding more characteristics including age, gender, and culture to the MLP neural network model. This research goes beyond simple emotion identification by creating a semantic multimodal non-observable system that links the context of the emotions detected by the MLP model with N-ary ontologies derived from semantic descriptions of events.

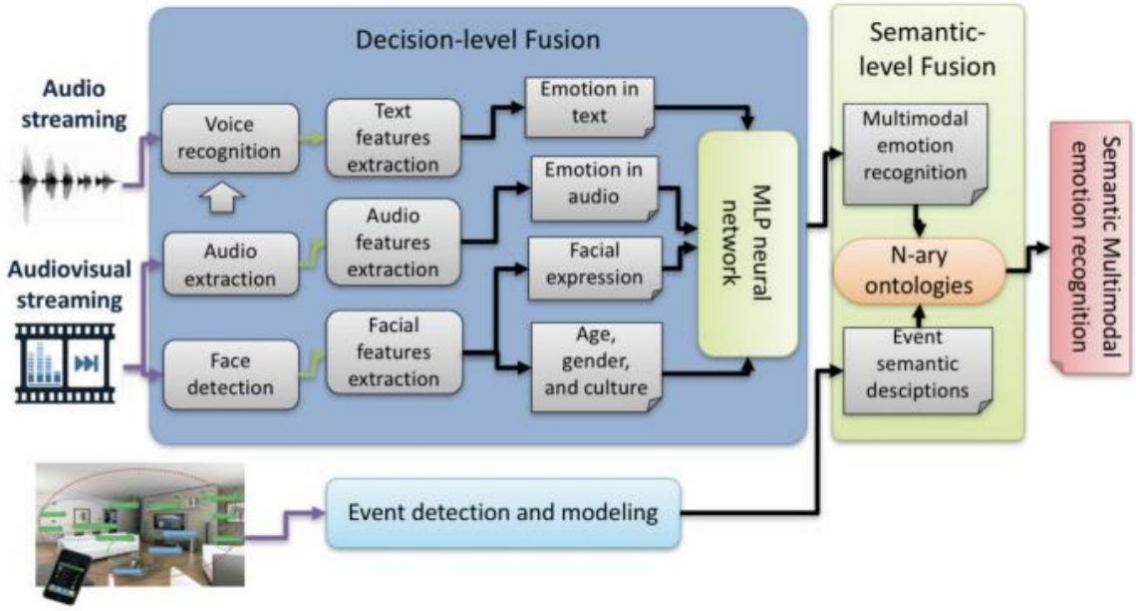


Figure 3.5: Multimodal Emotion Recognition architecture in an intelligible robotics system

3.3 EXPERIMENTAL SETUP

Python was used on a Windows 10 operating system with a configure ration that included 4 GB of RAM to perform the experimental setup for the suggested strategy. The main data source for the studies was the DEAP dataset. With this dataset selection, we can be certain that we have a rich and uniform basis on which to measure the scheme's efficacy.

Figure 3.6 shows that the proposed approach begins with data gathering, in this case images of faces displaying various emotions. As part of its feature extraction procedure, audio

processing and facial expression recognition are used to gather pertinent data. There are many subprocesses inside the VGG-16 CNN method, such as the pooling layer, natural language processing, and regularization. Through the use of the suggested approach, the model is able to identify mental health issues.

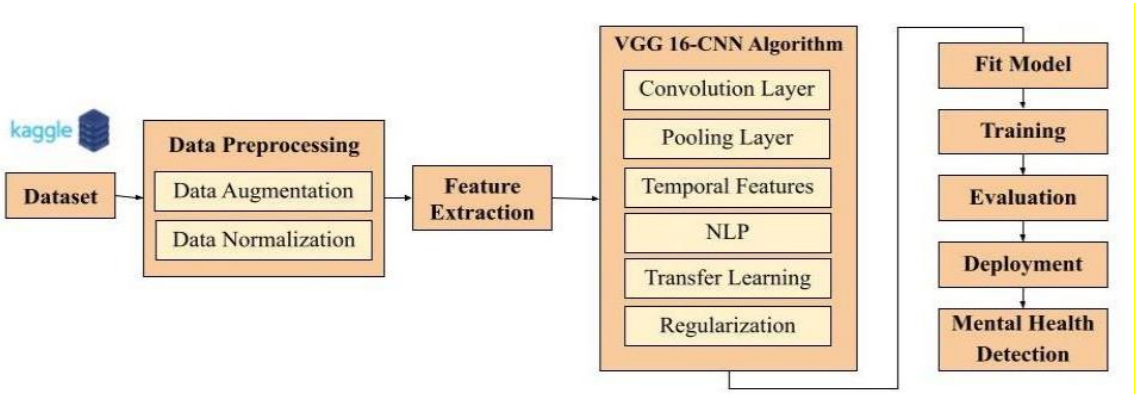


Figure 3.6: System architecture

3.4 DETAILS OF DEAP DATASET

The first step in the proposed model images of face expressions conveying various emotions. The process of feature extraction makes use of audio processing and facial expression detection to glean useful information. The dataset is gathered from Kaggle and it consists of 19 classes which contain 8 different types of emotions of different people. Images are categorized based on different kind of facial expression expressions like disgust, fear, happy, neutral, sad, surprised, anger and contempt. A dataset which is used for sentiment analysis purposes is likely to contain a variety of attributes or features related to sentiment, such as emotion-based images and their corresponding textual descriptions or annotations.

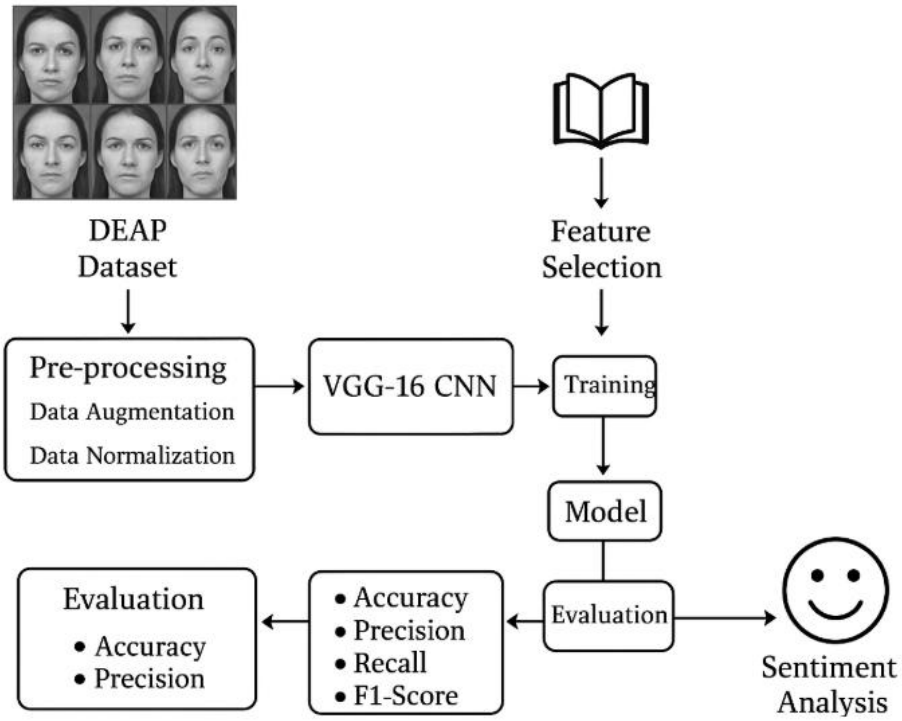


Figure 3.7: Deep Dataset Emotion Analysis Workflow Diagram

Examples of data pre-processing procedures include data augmentation and data normalization. Pre-processing and data preparation for testing and training may be handled via TensorFlow's Image Data Generator. Sentiment analysis makes use of the VGG-16 CNN model, which has its feature set enhanced by feature engineering and selection. When preparing picture data for deep learning tasks, it is common practice to utilize the 'picture Data Generator' tool, which is part of the TensorFlow and Kera's libraries. When it comes to picture classification, using a CNN architecture that excels at collecting spatial data is paramount. The 'validation split' option provides the main splitting method. The training data makes up 80% of this dataset, while the validation data accounts for the remaining 20%.

Training, validation, and adjusting hyperparameters are the subsequent processes. To measure how well a model works, researchers employ evaluation measures including recall, accuracy, precision, and F1-score matrices. It is possible to have empathetic conversations by studying facial expressions, which may aid with real-time emotion prediction. Then, current systems may be enhanced by incorporating future research topics related to mental healthcare and content customization.

The majority of the collection consists of participant assessments and publicly available online self-assessment scores. It provides a goldmine of information for research on emotion classification. Numerous physiological recordings, such as electroencephalogram (EEG) readings and facial film, are part of the collection with the aim of emotion analysis.

In it, participants rate 40 music videos while simultaneously collecting physiological data like EEG signals. Another twenty-two people had their likenesses recorded. Using this dataset as a springboard, researchers may investigate emotion detection across several

modalities.

3.5 DATA COLLECTION

Because data is the lifeblood of machine learning algorithms, picking the right dataset was a critical step in building the system. The job may have been made much more difficult had the wrong dataset been chosen. In order to conduct this study, we will be using structured databases to collect trend datasets. Choosing For instance, face datasets would have unnecessarily complicated the face detection method, even if the main objective of this study is not facial identification.

The image quality of various facial expression databases is varied, and access to many high-quality datasets is permission-based. Thankfully, acquiring permission is often simple, since the majority of as a formality, dataset owners merely need users to fill out a form. Following a thorough analysis, the trend VGG16 and ResNet50 models were chosen for this study's loading onto DCNN. These models provided labeled photos with a noise-free backdrop and the key face expressions.

3.6 IMAGE PREPROCESSING

In the past, preprocessing—the act of transforming raw data before feeding it into a deep learning algorithm or neural network—was considered a crucial step. In contrast, convolutional neural networks (CNNs) are not as dependent on specific dataset preparation, according to the claim. Reason being, convolutional neural networks (CNNs) may train themselves to recognize important characteristics only by looking at the raw pixel values of pictures. By reducing the need for considerable human preparation, CNNs streamline the workflow in image-related jobs by automatically extracting hierarchical characteristics. Several typical processes are involved in image preparation for Deep Convolutional Neural Networks (DCNNs), including Google Net (Inception), ResNet50, and VGG16. The goal of these procedures is to improve the model's capacity to learn relevant picture attributes by standardizing the input data.

Image preparation often involves the following steps: Make sure the input picture dimensions are the same as what the neural network design anticipates. Image resizing to meet model specifications is essential. For instance, VGG16 requires 48x48 pixels, ResNet50 224x224, and Google Net 224x224. To make sure the model is getting consistent input, normalize the pixel values to a common scale. To make sure that all of

the pixel values fall inside the interval [0, 1], divide them all by 255. The data may be centering on zero by subtracting the average pixel value from the dataset. This is a popular approach to get the same outcomes as the original, pre-trained models. Subtract the mean pixel value from the dataset for every pixel in the input image. Add adjustments like rotations, flips, and shifts to the training data to improve the model's ability to generalize. While training, but not while validating or testing, apply random alterations to the training pictures.

3.7 FACE DETECTION

In the processing pipeline, face detection was the first and most important phase. We could not go further with the analysis without first identifying the face, even though our photographs only include frontal facial expressions. It became much easier to locate the area of interest and extract key data when face detection was successful. Different face identification techniques were tested, one of which being OpenCV's Haar-cascades. In the end, we settled on a face detector that used the Dlib library's Oriented gradient histogram (HoG). A combination of Support Vector Machines (SVM) and HoG descriptors is needed for precise face recognition in images worked well. The photographs were shrunk and grayscaled during the preparation step to make sure they were all the same.

3.8 PRE TRAINED MODELS

An annual event is the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). that showcases various photo categorization models developed over the years. Building new models of deep convolutional neural networks would be substantially aided by implementing these training and design breakthroughs. After training on millions of images across several categories, models have found feature weights that work exceptionally well for certain jobs.

Since APIs often make them publically accessible, new models may integrate parts of existing pre-trained models. The components will either operate as a feature extraction process or use the weights of the identified features to formulate new interpretation problems. Keras is great since it's easy to use, has pre-trained CNNs available, and you can acquire model weights for free from their libraries. We are free to pick and select from the state-of-the-art models and use them to solve our problem. CNNs that use residual modules (ResNet50), inception modules (GoogLeNet), or repeating structures (VGG16) from the Keras applications are examples of appropriate CNNs. Images from

convolutional neural networks (CNNs) follow a preset algorithm. Images may be converted to a certain pixel input shape or even in grayscale format for your own model's training or model weight calculations. This is useful for importing models from their database and building a new baseline model. You will have to adapt your own image format to what these models need in order to use the results that these architectures have previously obtained in terms of feature recognition or model weight calculations.

Models from the deep learning domain can only be used in Keras applications when pre-trained weights are available. Feature extraction, prediction, and fine-tuning are just a few of the many uses for Keras models. This article will explore the topic of Keras applications in great detail. In an effort to understand it, we will utilize the following subtopics: "what are keras applications?", "how to use models from Keras. applications?", "keras applications model," and "a conclusion about the same."

We have models available for Keras.

Table 3.3: Keras Models Available for Use.

Model (MB)	Size	Top-1 Accuracy	Top-5 Accuracy	Parameters	Depth
Xception	88	79.0%	94.5%	22.9M	81
VGG16	528	71.3%	90.1%	138.4M	16
VGG19	549	71.3%	90.0%	143.7M	19
ResNet50	98	74.9%	92.1%	25.6M	107
ResNet50V2	98	76.0%	93.0%	25.6M	103
ResNet101	171	76.4%	92.8%	44.7M	209
ResNet101V2	171	77.2%	93.8%	44.7M	205
ResNet152	232	76.6%	93.1%	60.4M	311
ResNet152V2	232	78.0%	94.2%	60.4M	307
InceptionV3	92	77.9%	93.7%	23.9M	189
InceptionResNetV2	215	80.3%	95.3%	55.9M	449
Mobile Net	16	70.4%	89.5%	4.3M	55
MobileNetV2	14	71.3%	90.1%	3.5M	105
DenseNet121	33	75.0%	92.3%	8.1M	242
DenseNet169	57	76.2%	93.2%	14.3M	338
DenseNet201	80	77.3%	93.6%	20.2M	402
NASNet Mobile	23	74.4%	91.9%	5.3M	389
NASNet Large	343	82.5%	96.0%	88.9M	533
EïcientNetV2B0	29	78.7%	94.3%	7.2M	-

Keras applications make use for prediction, feature extraction, and fine-tuning deep

learning models. When the model is constructed, all required weights are immediately added to the /keras/models' folder. The image data format and all the parameters shown in the Keras file, which is found under keras/Keras.json, are used to build the models. when the Keras applications are instantiated. Imagine, for the sake of argument, that channels_last is the value specified to image_data_format. When the data format follows the TensorFlow standard, which includes height, breadth, and depth, the model loads automatically from this repository.

You may incorporate these pre-trained convolutional neural networks (CNNs) into your own computer vision code for picture and emotion recognition jobs recognition via a process called transfer learning.

3.9 IMPLEMENTATION TOOLS

Choosing the proper high-level programming language to develop the system was difficult since there are now over 2000 of them accessible. Fortunately, this problem was made simpler by the project requirements. The programming language utilized during the load trend model phase has to be user-friendly, realistic, and compatible with the GPU libraries since neural networks were employed in the project. It was also crucial that there were sufficient documentation and community backing to address development cycle difficulties promptly.

The neural network implementation and GUI design should be done in a straightforward and easy-to-understand programming language. Programmers classify real-time visuals; therefore, the language must be quick enough to process this kind of data and the classification procedure. Due of these problems, the list was significantly shortened. The most widely used languages for machine learning include R, C++, Python, and MATLAB, among many others. R's reputation as a powerful language is unaffected by its challenging learning curve. However, learning C++ is infamously hard. and much more so to optimize, despite its reputation as the top option when performance is paramount. As a stand-in, Mat Lab is fast, simple to learn, and performs adequately. The development cycle for Python is lengthier than that of other languages, and a license is required to use the core product and parts of its modules, but the language is simple to learn and contains several packages that allow for transparent GPU use. Based on the selected Library and Tools, the next section details the components used to build the system.

3.10 DEVELOPMENT

The development phase is as critical, but it carries the risk of underestimating development time or producing tough, poorly written code. So, to avoid these dangers, a method based on modular design was used. By breaking down a system into smaller components, modular design streamlines software development. This system's basic idea is that every part has a specific job, and that changing one module has little to no effect on the others.

- At this stage, we divided the training and application parts. The training part was designed using the following modules:
- Upon receiving The DCNN module constructs the DCNN model using the labels and features. The input module is in charge of extracting these components.
- For The following modules were created for the application portion:
- The main element, the GUI module. You may see the input data, facial expression classifications, and the probability of each expression in this module.
- Similar to the picture is processed by the input processing module in training. Input data is read from the source picture, and the image is processed appropriately.
- The DCNN application module utilizes the trained DCNNs to provide the classification and probability for every phrase after receiving the pre-process vector.

3.11 EVALUATION METHODS

If you want to know how useful and effective a model is, you have to look at it. To find out what class an instance belongs to We use assessment measures like F1-score, recall, accuracy, and others in computational problems like detection and classification. precision. In order to calculate these values, the detection metrics were used. A model's performance for each class may be learned from the detection metrics. Using the value of the confusion matrix, the following metrics were created:

3.11.1 Confusion Matrix

Classification models bring forth a number of outputs that are categorized. Most error metrics can tell us how off our model is generally, but they can't pinpoint where it went wrong. A traditional measure of accuracy would not reveal if the model over-or under classifies certain groups.

It is common practice to use the confusion matrix when evaluating machine learning classification models. To obtain a full picture of how well the model worked, this matrix lets you compare the expected and actual values of the target characteristic.

Table 3.4 The confusion the matrix is 2x2., and it is used to classify binary data.

Confusion Matrix		Actual	
		Emotion Detected Positive (1)	Emotion Not Detected Negative (0)
Predicted	Emotion Exist Positive (1)	TP [1,1] True Positive	FP [1,0] False Positive
	Emotion Not Exist Negative (0)	FN [0,1] False Negative	TN [0,0] True Negative

Regarding binary classification model outputs, Table 3.2 provides a confusion matrix. The two categories that the model is supposed to differentiate between are "Positive" and "Negative" inside this matrix.

- Verified Assured:
- When both the actual and anticipated values are 1, we say that there are such cases.
- Verifiably negative:
- True negatives are situations in which both the actual and expected values are zero.
- Erroneous Positive:
- The illustration of a false positive would be a projected value of 1 while the actual value is 0.
- Negative Rejection:

In a false negative, the expected value is zero even if the actual value is one

.

- **Accuracy Rate**

Accuracy refers to how well the instances fit the categorization criteria. The frequency with which the model predicts the class accurately, that is, with a normal face image and an appropriate emotion picture, is addressed in the following explanation. It is calculated using this formula. In most cases, it is a metric for the classifier's accuracy rate. Accuracy is a number between zero and one.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision**

The probability that a positive value prediction would come true is shown by it, and it is a measure of how well all classes have predicted; for instance, determining the possibility that an image has an illness. Accuracy may be determined using the following formula:

$$Precision = \frac{TP}{TP + FP}$$

- **Recall**

Finding Compared to the total number of positive tests; the percentage of properly detected positive samples provides the recall. The accuracy with which the model can detect positive occurrences is quantified by this statistic. An increase in recall suggests that positive samples are being detected with more frequency. The recall measure, when applied to the task of transaction prediction, represents the percentage of positive predictions that are really realized.

$$Recall = \frac{TP}{TP + FN}$$

- **F1-Score**

One of the measures becomes zero when the lowest possible F1-score is zero. The harmonic means of the accuracy and recall scores is known as the F1-score. suggests a great degree of precision or memory. This is how the computation's outcome appears. The harmonic means of the recall and F1 score is accuracy scores, in most cases.

$$F1\ Score = 2 * \frac{Precision * Recall}{recision + Recall}$$

And let's say the statistics show that the classes are really different from one another. While a model may consistently outperform other models when it comes to predicting the majority class—the one with the most data instances—it will struggle to do the same when it comes to predicting the minority classes. Confusion matrices seem to be valuable in such instances.

3.12 MULTIMODAL EMOTION RECOGNITION FOR HUMAN- COMPUTER INTERACTION

- **Data acquisition**

Gathering raw data from various sensors is what data acquisition is all about, depending on whether the data is to be integrated based on discrete or continuous emotional states and the modality that is to be used. Important for both the design and execution of automated emotion identification, this stage requires data that represents almost every conceivable real-world scenario. Many of these studies rely on data that was either created or obtained from publicly available datasets or from researchers' own databases. Since this research does not aim to go into the specifics of dataset design setup, we will instead provide a high-level overview of the multimodal emotional datasets that have previously been published.

There are three categories of these datasets based on the ease with which researchers can access them: those with full access, which are free to download and specifically made for researchers, those with partial access, which are available for free download upon the signing of an End User License Agreement, and ultimately, the commercial datasets, which require payment. But the greatest dataset is the one you create yourself as it should include everything you need for the project. More time and specific settings are required, which is the only drawback.

The modalities that are of interest might be visual, auditory, textual, physiological, or keystroke-based. Any number of modes, from one (unimodal) to many (multi-modal), may be present. Task requirements and available resources dictate the mode of operation. Facial expression datasets are the most common because recorders for them are readily

available (cameras, webcams, etc.), whereas bio-physiological based datasets are the least common due to the difficulty in accessing the recording material and the obstructions that can be caused by need sensors on the body.

Because issues like sensor failure may introduce mistakes and noise into the collected data, the quality of the recorded material is important. While a reduced data set could make training easier, it severely limits the test's ability to generalize. A full comprehension of these facts is required in order to continue effectively. The collected data may include unnecessary information, inaccurate or irrelevant information, or missing values. As a result, during data pre-processing, they must be prepared for further processing.

- **Data Pre-Processing**

In order to get the data ready for subsequent processing so it may be used as intended, a procedure called data pre-processing is necessary. Data cleansing, integration, reduction, and transformation are the four main phases of data pre-processing. If you want to know more about the procedures mentioned above, you may read Pattern Recognition by Theodoridis and Kourtroumbas and Han et al.

During data pre-processing, the primary focus is on:

We have divided these datasets into three categories according on the context: induced, spontaneous, and performed. actors were recorded. Samples in which the individuals' emotions are predetermined are called acted datasets. When doing research, they are the most accessible and convenient to get. One drawback of such datasets is that capability; in reality, the trained samples' preset conditions aren't robust enough to handle real-world scenarios. To address certain stereotyped expressions in the raw data, however, perception tests may be used to filter out exaggerations and abnormal behavior.

The actors in the induced emotion datasets are put in scenarios where they are expected to organically display different emotions based on the circumstances. When compared to the performed, they're superior since they let the subject show their feelings naturally, without the designer having to intervene. Nevertheless, they still have limitations that cause humans to inhibit or exaggerate their emotions because of the recording process's awareness.

Due to concerns about the players' privacy and security, natural (spontaneous) datasets are notoriously difficult to get, while being superior from a practical standpoint. Media such as YouTube, Facebook, Twitter, and other social and multimedia websites, as well as television shows and live news broadcasts, are the primary sources for these in the literature. Their extensive recording capabilities and difficulty in appropriate annotation are their main drawbacks.

- **Features Extraction**

For the data gathered so far, each example falls into one of many classes or categories, and although all of the samples share the same characteristics or qualities, their values vary. Features, input variables, or characteristics are the names given to these variables. Depending on the modality, they may be continuous, binary, or categorical, and serve as a representation of the data (Table 3.3). Identifying the modality-specific feature is what feature extraction is all about. Table 3.3 shows that fundamental features, which might include additional intermediate characteristics, are often defined during feature extraction in multimodal emotion identification. A vector feature of the modality will be composed of many feature values. For the purpose of emotion recognition, Table 3.3 summarizes several features that have been extracted:

Table 3.5: Feature extraction and modalities

Modality	Feature		
	Basic	Intermediate	
Face ex- pression	eyes, eye- brows, nose, mouth	Action Units	
Speech	linguistic	Words, multiword, phrases, sentences, documents	
	Paralinguistic	Pitch, bandwidth, duration, pro- sodic, voice quality, Mel frequency Cepstral coefficients (MFCC)	
Body	Head gestures	Head position	
		Head movement	
	Hand Gestures	Shape	
		Motion	
	Body motion	Spinal column	Neck, chest and Abdomen
		DOF body	Symmetrical arms
		Body center mass	Movement of body center of mass
		Joints	Degree of joint rotation
Physio- logic	Hearth, Brain, Limbs, Blood	Electrocardiogram (ECG); Electro dermal activity (EDA); Electromyogram (EMG)	

An unanswered question in the field of emotional states research is how to precisely identify the most important quality to exhibit. Voice, body language, hand gestures, facial expressions, and physiological cues are the most often used modalities in research; however, this largely depends on the accessibility and availability of the datasets. There are two parts to the features extraction process: building the features and selecting them.

In feature creation, the available measurements and domain criteria are used to determine the appropriate data representation. In addition to automated feature development techniques, the features may also be produced manually. There have been several recent suggestions for automated feature extraction tools. These tools would be useful for extracting certain features. Table 3.6 provides a summary of certain automated features extraction techniques.

Table 3.6: Feature extraction techniques and algorithms are included

Toolkit	Modality	Feature extracted/ Function	Brief description
PRAAT	Audio	Duration, F0, Range, Movement, Slope, Energy features	PRAAT (a system for doing phonetics)
FEELTRACE	Audio	Labelling	FEELTRACE, allow- ing the emotional dynamics of speech episodes to be examined.
Open EAR	Audio	Signal Energy, Loudness, Mel- /Bark-/Octave- Spectra, MFCC, PLP-CC	Open EAR provides efficient (audio) feature ex- traction.
Open SMILE	Audio	Signal Energy, Loudness, Formants, Mel-/Bark- /Octave- Spectra, MFCC, PLP-CC, Pitch, Voice quality (Jitter, Shimmer), LPC, Line Spectral, Pairs (LSP), Spectral, Shape de- scription	It is an open-source toolkit, for feature extraction in ma- chine learning and data mining by re- searchers in voice processing, vision signal processing and Music Information Retrieval.
Eyes Web	Body	Quantity of motion, cue, Contraction index of the body, velocity, Acceleration, fluidity of the hand's, barycentre	Open software for extended Multimodal Interaction. University of Genoa - Italy
Lux and FSDK 1.7	Face	Action Units	Facial recognition software
ANVIL	Audio	Annotation tool in a multi-modal dialogue	Free for research purposes.
AAM	Face	Texture and shape	It uses a statistical approach
GAVAM	Face	Face component detection	Constrained local models

The majority of the tools listed in Table 3.6 are open source and available for download online. The most widely used operating systems, such as Linux, Mac OS X, and Windows, are largely compatible with them. After building features, the next step is to choose the most important and helpful ones.

Choosing features that are more instructive and applicable to the current job is the primary goal of features selection. As a rule, feature building might result in hundreds of features, which slows down training and necessitates a significant quantity of storage. To get over these problems, feature selection employs a broad data reduction strategy by reducing the amount of unnecessary information to a manageable minimum. Data comprehension and performance enhancement are both aided by features selection. The key is to choose characteristics that have a limited variation within the same class and a wide distance across classes. We suggest that readers consult Theodoridis and Kourtroumbas for more

information.

3.13 INFORMATION FUSION

Decision (or semantic) The information is merged at three different levels: level fusion, feature level fusion, and data level fusion. However, feature and decision level fusions are heavily used in multimodal emotion detection. These combinations stages are briefly described below.

- **Integration at the Data Level.**

Raw data is the only thing that undergoes data-level fusion. For data of the same kind, such as many cameras taking pictures of the same scene from various angles, it is best to use a comparable kit. Although noise contaminates the raw data, the data-fusion level is theoretically higher, resulting in enhanced accuracy due to less loss of fundamental information. More precise data level fusion requires costly processing.

- **Fusion of Features.**

The process of fusion involves merging information from several modalities into a single vector that may be used to train a classifier. In order to achieve simple generalization, feature fusion works best when the modalities are strongly connected and coordinated. This is the point when it becomes necessary to gather a substantial quantity of training samples because of the fact that the input features' dimensionality will have increased.

- **Fusion at the Level of Decisions.**

At the pinnacle of combination is the decision-level fusion. The decision-making process begins with training separate classifiers for each modality, and then the result is aggregated and fed into a single classifier. This has to do with combining asyn-chrous and temporally related methods. Because each modality is learned independently and integration does not need retraining, it is more preferred in practice than the preceding ones. Methods for reaching a verdict in published works range from decision trees and Fisher Discrimant Analysis to the weighted sum rule.

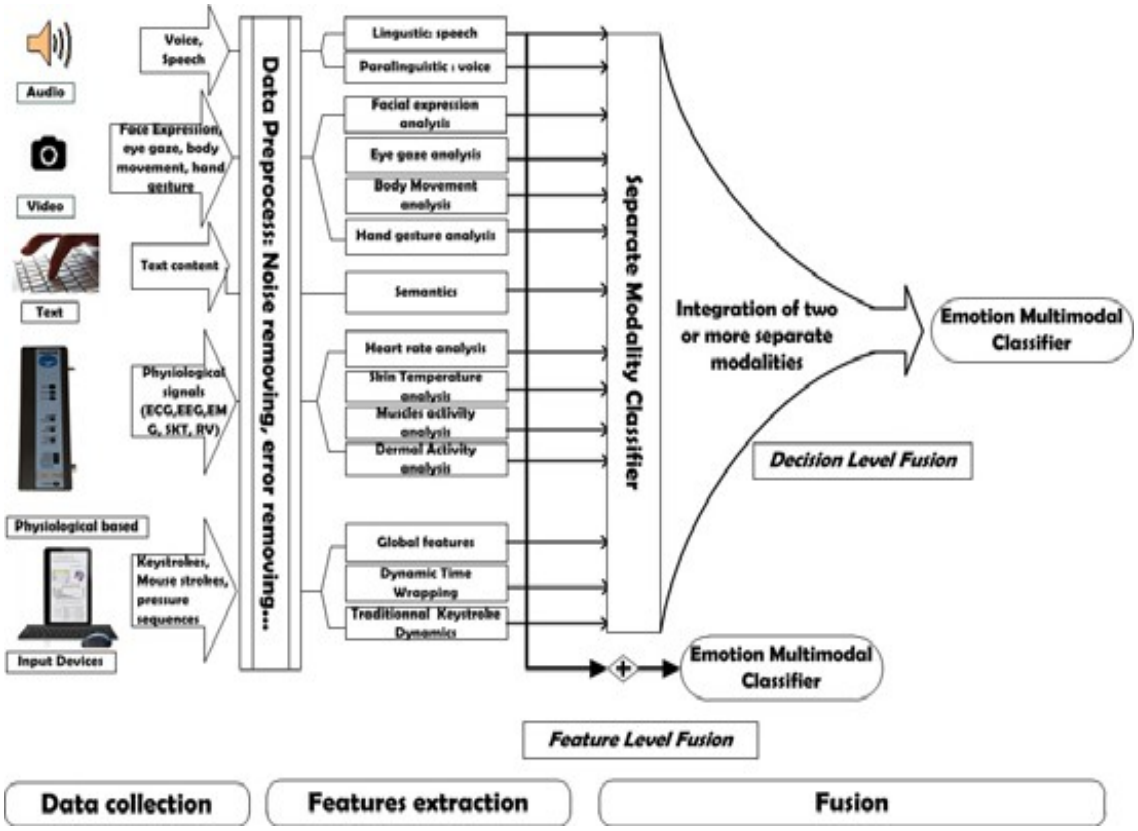


Figure. 3.8: Multimodal Emotion Recognition Design System conception

Since There isn't yet a widely recognized technique for determining the effectiveness of data fusion algorithms, evaluating them poses the greatest challenge. Research should concentrate on developing systems that can educate, and on enhancing knowledge, abilities, and plans via experience; in most learning scenarios, the solution is tested under settings that are comparable to training. Janssen et al.'s study demonstrates how to establish a benchmark dataset by showing that robots surpassed humans in emotion identification.

Although The identification of multimodal emotions is still in its infancy, especially in some domains like emotion recognition input devices, as Picard points out, a partial answer might still be useful in a challenging circumstance. The development of more efficient machine learning algorithms indicates that the time has come for practical, real-time applications.

3.14 FACIAL EMOTION RECOGNITION USING MULTIMODAL SIGNALS

When it comes to decision-making and social connections, among other areas of human life, emotions play a major role. In many fields, including education, healthcare, and customer service, multimodal emotion detection is crucial. In order to achieve emotion

identification, this research employs multimodal signals, which include physiological data, face photos, and EEG. First, to eliminate any undesired artifacts, the facial image is pre-processed. After that, features are extracted from the processed picture using LDTOOP, a combination of LOOP and LDTP. Similarly, features like wavelet coefficients and spectrum parameters including PSD, spectral flux, tonal power ratio, spectral skewness, and spectral centroids are used to get the EEG and physiological signals ready for analysis. extracted. Afterwards, a DBN classifier processes the characteristics obtained from the face picture.

3.14.1 Gathering of input video

An important aspect of emotional computing systems is emotion recognition. A person's emotional state is multi-faceted, including not just their thoughts and actions but also their physiological responses to external stimuli. Physiological signals, video, and electroencephalogram (EEG) data are common components of facial emotion detection systems. When training and testing face emotion detection algorithms, one popular dataset is the DEAP dataset.

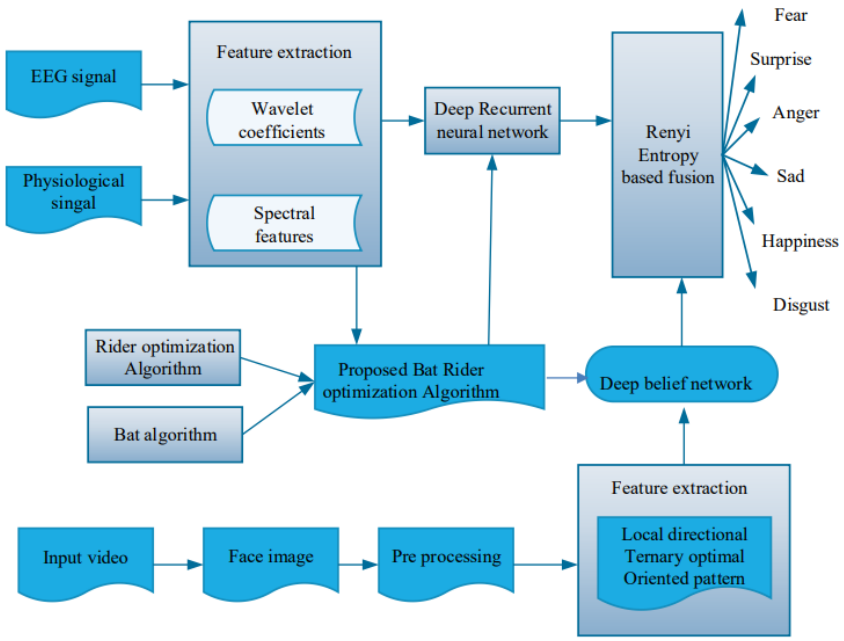


Figure 3.9: Proposed BROA-based Deep NN schematic view

3.14.1.1 Image extraction for the face

Here we'll assume a dataset A and use the variable to indicate the number of movies

included inside it.

$$H = \{A_1, A_2 \dots A_i \dots . A_n\} \qquad (3.1)$$

Here, H stands for the dataset and n for the number of movies. Face shots are included in each film from different angles. In dataset H, the ith movie contains facial pictures, which may be represented as follows:

$$A_i = \{X_1, X_2 \dots X_i \dots . X_n\} \qquad (3.2)$$

The ith facial picture from the ith video is represented by A in this context. There is a unique location for each individual's face picture, and each face image belongs to a certain person. The pre-processed result is the result of taking the face picture from the dataset and running it through a pre-processing step.

3.14.1.2 Image pre-processing for faces

To ensure that the input face picture is of high enough quality and suitable for further analysis, the pre-processing module is an essential part of face image analysis. The main goal of the several processes that make up this module is to remove noise and external artifacts. After importing the face picture into the pre-processing module, the image undergoes noise reduction.

- **Noise Removal:** Electronic interference, changes in illumination, and imaging equipment flaws are among the many potential sources of picture noise. The reliability of later analyses and feature extractions may suffer because of this noise. Noise removal methods are used to clean up the picture and improve its quality in order to solve this problem. Median filtering, Gaussian filtering, and wavelet denoising are some of the most used techniques for noise reduction. These techniques aim to eliminate noise from the picture without affecting the essential face details.

- **Artifact Removal:** In addition to noise, the face picture may include extrinsic artifacts such occlusions, specular reflections, or shadows. These distortions may make it harder to provide a precise diagnosis by masking facial characteristics. These artifacts are detected and removed by the pre-processing module via the use of picture in-painting methods to fill in missing or obscured sections of the face or by recognizing and removing regions with uneven intensity patterns.

- **Normalizing Lighting:** Changes in the surrounding light may have a dramatic effect on how a face looks in a photo. To make sure the analysis doesn't alter as the lighting changes, you may use illumination normalization methods. To facilitate the extraction of consistent characteristics, these methods seek to uniformly illuminate the face. Lighting normalization approaches include histogram equalization, adaptive histogram equalization, and methods based on Retinex.

- **Image registration:** This method brings together disparate photographs of an individual's face and aligns them with a preset set of coordinates. When working with many face pictures for comparison or tracking, this is a necessary step. The spatial arrangement of facial markers, including the mouth, nose, and eyes, may be ensured. remains consistent across multiple images using techniques like affine or non-rigid image registration.

- **Face Detection and Alignment:** Face identification and alignment are additional tasks that may be included in the pre-processing module. Face recognition finds the face in the picture, and alignment makes sure it's centered and facing the right way. Consistent feature extraction relies on this stage. Common techniques for face identification and alignment include deep learning-based methods like Convolutional Neural Networks and Haar cascades.

- **Increasing the Clarity:** Depending on the use case, resolution enhancement methods may either make the picture sharper or more detailed. One approach is to use interpolation or super-resolution algorithms to upscale the picture. For scalability, this is not necessary. The final picture, M pre-processed, is cleaner, more uniform, and ready for further examination after the pre-processing step. Tasks like face recognition, emotion analysis, and facial expression classification rely on the feature extraction module, which takes this pre-processed picture as input. When it comes to robust and reliable facial picture analysis, the pre-processing module is very important.

For feature dimension reduction, Features from the pre-processed output must be extracted. In this instance, the built-in LDTOOP capability is used to extract features. Combining the Local Optimal Oriented Pattern (LOOP) with the Local Directional Ternary Pattern

(LDTP) results in form this characteristic.

Image processing and computer vision make use of texture descriptors like LOOP and LDTP for several uses, including object detection, picture categorization, and texture analysis. Using various approaches, these descriptors extract texture information from specific areas of images. Now we'll go over each one in detail:

When it comes to pattern identification and texture classification, computer vision and image analysis often turn to the "Local Optimal Oriented Pattern" (LOOP) as a feature extraction approach. Texture classification, object identification, and facial expression analysis are just a few of the many applications that greatly benefit from its ability to capture important spatial information and directional properties inside pictures.

Splitting a picture into smaller, more specific areas, also termed cells or patches, is the basic idea underlying LOOP. An orientation histogram is produced for each of these specific areas by the method. You can see the distribution of gradient orientations in each area in these histograms. Here, gradients denote the varying rates of intensity change throughout the picture. You can learn a lot about the local area's texture patterns and edge structures from the data in gradient orientations.

To extract complex structural and textural details from pictures, Professionals in computer vision and image processing use the latest -art Local Directional Ternary Pattern (LDTP) texture analysis method. By using a ternary encoding technique, LDTP expands upon the Local Binary Pattern (LBP) basis and improves its texture description and classification capabilities.

One way LBP works is by comparing the intensity values of nearby pixels to a center pixel in a local neighborhood. If the intensity values are equal or higher than the central pixels, then the adjacent pixels are given binary codes (1s and 0s). To illustrate this point, LDT gives three potential results for every comparison: more than, equal to, or less than the intensity value of the center pixel. A change from LBP's binary encoding to 1, 0, and -1 represents these results. Using this ternary encoding approach, LDT improves the depiction of local texture patterns. The Least Difference Time (LDT) algorithm generates a series of 1s, 0s, and -1s for every pixel in a picture by comparing it to its nearby pixels. A feature vector for a whole picture or a specific area is generated by building histograms of these

ternary patterns. The distribution of different geographical patterns is captured by these histograms, which together provide a feature vector that provides structural details and a description of the picture.

The enhanced rotation invariance of LDT is a significant benefit over conventional LBP. Because it takes the directional connections between pixels into account, LDT is more resistant to fluctuations produced by picture rotations. Robust texture analysis across rotated versions of the same texture is made possible because the relationships between pixel values in LDT patterns are retained when a picture is rotated.

When combined with machine learning techniques like SVMs or deep neural networks, LOOP and LDTP texture descriptors are a powerful tool for picture analysis. Which one is better relies on the nature of the work and the studied photos' textures.

We may think of the pre-processed image's intensity at pixel position lc, which is the intensity of a pixel inside a neighborhood of size [3x3]. Value of the loop for the pixel located at location (c_r, d_r) is able to represent as:

$$L_m(c_r, d_r) = \sum_{n=1}^{7} \lambda(l_c - l_i).2^{\beta n} \qquad (3.3)$$

$$\lambda(c) = \{1; c \geq 0\ otherwise\ \} \qquad (3.4)$$

Rotation invariance is introduced to the basic formulation by using the LOOP descriptor. The elimination of the LDP's arbitrary parameter assignment is its intended use. The LDTP combines the Gaussian and Frei-Chen compass masks. By reducing noise disturbances, Gaussian masks make models more resilient to changes in illumination, while the Frei-Chen mask brings out the embedded structural details. LDTP encodes the texture of the picture by using these masks to calculate edge responses in eight directions.

In LDTP, contrast information is encoded by comparing periphery pixels with core pixels. The LTP model serves as the foundation for developing LDTP code. The method successfully sorts local primitives and retrieves improved data because edge responses are more robust against noise and light than simple intensity values. Variations in texture and intensity of the structure are specified by the center pixel within a [3x3] square region. An evaluation of the center pixel's relative edge response is carried out by the LDTP operator. With the center pixel denoted as and the outermost pixel, l_e, where i is an integer between

0 and 7, l_d, the average local gray level of the neighborhood consisting of a [3x3] square is represented as:

$$\mu = \frac{1}{9}\left[l_d + \sum_{i=0}^{7} \quad l_e\right] \qquad (3.5)$$

In addition, LDTP code is generated using the LTP principle. The LTP framework breaks down any ternary pattern into its positive and negative parts. To create LDTP code, this division is used. In this case, the LDTP upper and lower boundaries are defined by the kernel function.

$$LDTP_{lower} = 2^8\gamma(\mu - l_d, EG_d) + \sum_{i=0}^{7} \quad 2\gamma^e(l_e - l_d, EG_e) \qquad (3.6)$$

$$LDTP_{lower} = 2^8\gamma^*(\mu - l_d, EG_d) + \sum_{i=0}^{7} \quad 2\gamma^{*e}(l_e - l_d, EG_e) \qquad (3.7)$$

In cases when, y^* and y is presented in,

$$\gamma(c,d) = \{1; c \geq 0 and d \geq 0\ 0; otherwise \qquad (3.8)$$

$$\gamma^*(c,d) = \{1; c \leq 0 and d \leq 0\ 0; otherwise \qquad (3.9)$$

An integrated feature is generated by merging the histograms obtained from the examination of the upper and lower LDTP scales. The g-shaped feature, with dimensions of [1 × 10] correspondingly, is the outcome of integrating this fused feature with the LOOP descriptor.

3.14.2 DBN-based facial emotion recognition

In order to categorize emotions from face images, the DBN classifier is used. The DBN classifier uses these characteristics as inputs to recognize face emotions. With connections created between visible and buried neurons, the DBN architecture consists of two RBM layers followed by one MLP layer. Importantly, the next layer takes as input the output of the previous layer.

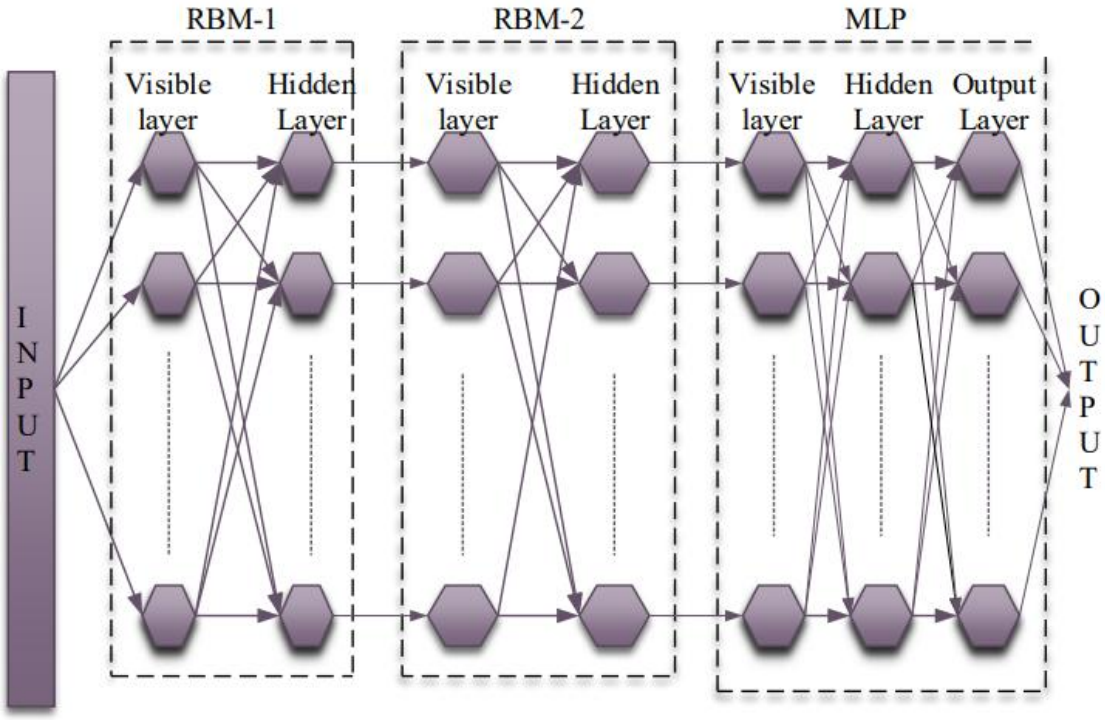


Figure 3.10: Recognizing emotions with DBN

Classification is only one of several machine learning applications with a certain architecture known as a DBN with two RBM layers and one MLP layer. Here is a detailed breakdown of the structure of this DBN:

The raw data, which might be photos, text, or anything else, is represented by the input layer. An attribute or dimension of the data is associated with each node in the layer of input. The first RBM layer receives the data from the input layer.

The visible and concealed units make up The DBN's first layer is the first RBM layer. Data from the input layer is used by the visible units in this layer. The RBM's primary goal in this case is unsupervised feature learning. It learns to identify significant characteristics in the input data and adjusts the weights and biases during training to reduce the reconstruction error between the visible and hidden layers. The first RBM layer starts extracting valuable features from the input data as soon as training is finished.

The second RBM layer creates a hierarchical representation of features by using the hidden units from the first RBM layer as the visible units. The feature representations learnt by the first layer are further refined by the second RBM layer. It follows the same

unsupervised training procedure as the original RBM and uses both visible and hidden units.

Adding a single Multilayer Perceptron (MLP) layer for classification is the last step after pre-training the two RBM layers. One layer of It is the multi-layer perceptron (MLP) layer of feed-forward neural networks. It uses the feature representations that the RBM layers have learned to do supervised fine-tuning for classification tasks. The number of classes in the classification problem, expressed in neurons, determines the size of the MLP's output layer. Typically, a SoftMax activation function is used to the output layer to create class probabilities. Using supervised learning algorithms like backpropagation and gradient descent, together with labeled data, the MLP's biases and weights are fine-tuned during training.

The DBN may be used for classification jobs when it has been trained to completion. To generate predictions from fresh, unlabeled data, the input is first sent through two RBM layers to The MLP layer is accessed after features have been extracted. Consequently, the output class is the one with the greatest probability prediction.

Using a single MLP layer in conjunction with two RBM layers allows DBN to integrate supervised classification with unsupervised feature learning. Several machine-learning tasks, particularly those that involve complex data representations, have effectively used this design. Using the DBN classifier has numerous advantages, such as its rapid training and ability to work with tiny datasets. Figure 3.10 depicts the arrangement of the DBN.

In RBM layer 1, Feature M generates the hidden layer's output and acts as the visible layer's input.

$$X^1 = \{X_1^1, X_2^1, \dots\dots\dots X_I^1 \dots . . X_N^1\} \qquad 1 \le i \le n \quad (3.10)$$

Given this setting, X^1 means the hidden neuron with the index i, and n stands for the total count of hidden neurons. There is a bias associated with every neuron in both the visible and buried levels. The following are the biases of the exposed and concealed layers:

$$R^1 = \{R_1^1, R_2^1, \dots\dots\dots R_u^1 \dots . . R_a^1\} \qquad (3.11)$$

$$S^1 = \{S_1^1, S_2^1, \dots\dots\dots S_i^1 \dots . . S_n^1\} \qquad (3.12)$$

Here we have it,R_a^1 is the bias that corresponds to the u[th] observable neuron, and S^1 shows how the nth hidden neuron is biased. Additionally, the weight that is initialized for RBM

layer-1 is represented as:

$$U^1 = \{U_{vi}^1\}, 1 \leq v \leq a; 1 \leq i \leq n \qquad (3.13)$$

In this context, U_{vi}^1 stands for the association between the V[th] visible and the ith hidden neuron in terms of weight. At RBM layer-1, the calculated result is shown as:

$$X_i^1 = \gamma\left[S_i^i + \Sigma M_v^1 U_{vi}^1\right] \qquad (3.14)$$

In this case, the activation function is symbolized by the symbol y. Layer 2's visible neuron count is equal to layer 1's hidden neuron count. However, R^2 and S^2 represent the biases connected to RBM layer-2's hidden and transparent layers, respectively.

This leads us to the following weight factor for RBM layer-2:

$$U^2 = \{U_{ii}^2\}; 1 \leq i \leq n \qquad (3.15)$$

This is at, U_{ii}^1 stands for the relative importance of the i[th] hidden neuron relative to the i[th] visible neuron.

So, the calculated result at RBM layer-2 is given by:

$$X_i^2 = \gamma[S_i^2 + \Sigma_v M_v^2 U_{ii}^2] \forall M_v^2 = X_i^1 \qquad (3.16)$$

Given this setting, R_i^2 shows the bias linked to the hidden neuron at the i-th position. Nevertheless, the MLP is fed data in the following format:

$$Y = \{Y_1, Y_2 \ldots Y_i \ldots . Y_n\} = \{Y_i\}; 1 \leq i \leq n \qquad (3.17)$$

In this case, the total number of input layer neurons is denoted by n. The following is the definition of the weight that connects the input and hidden layers:

$$U^v = \{U_{iz}^v\}; 1 \leq z \leq \beta \qquad (3.18)$$

Regarding this matter, U_{iz}^v is the weight that links the hidden neuron z[th] with the input i[th]. Here is how the hidden layer output is computed with weight and bias in mind:

$$P_z = \left[\sum_{i=1}^{n} U_{iz}^v * Y_i\right] \lambda_2 \forall Y_i = Z_i^2 \qquad (3.19)$$

This is at, λ_z stands for the hidden neuron's bias. Afterwards, the MLP layer receives its input from the RBM layer-2. The following is an illustration of the weight U^x, which connects the hidden and output layers:

$$U^x = \left\{U_{zo}^y\right\}; 1 \leq z \leq a; 1 \leq 0 \leq a \qquad (3.20)$$

Here is the formula for the output vector N_o, which makes use of the buried layer's weight

and output

$$N_o = \sum_{z=1}^{a} U_{zo}^{v} * O_z \qquad (3.21)$$

This is at, U_{zo}^{v} symbolizes the weight that links the v-th hidden neuron O-th to the output neuron, and O_z denotes the hidden layer's output.

i. Acquisition of physiological and EEG signals

In response to sudden shifts in mood, the brain quickly produces the EEG signal. Notable features for emotion identification are provided by these EEG data. Every person's EEG and physiological data are collected in order to build an emotion detection system. The physiological signal obtained from the ith individual in the video is represented by P, while the EEG signal corresponding to the ith person is represented by E I_i has the corresponding H label.

ii. Extraction of features from physiological and EEG signals

Electronic brain waves and physiological signals the emotion recognition procedure makes use of H to extract attributes. A number of characteristics, including wavelet coefficients, spectral properties including Signals E and E are processed to obtain PSD, tonal power ratio, spectral flux, spectral skewness, and spectral centroid. H.

- **A wavelet's coefficients**

Within a small local region, it establishes relationships. As a result, nearby components with increased values make up a large wavelet coefficient. The resulting wavelet coefficients have a size of [1 × 50] and are represented as d_1.

- **Spectra of variables**

The EEG and physiological data are used to determine several spectrum properties, such as PSD, spectral flux, spectral skewness, spectral centroid, and tonal power ratio. PSD is used to determine the input signal's power distribution within a certain frequency range. It has dimensions of [1 × 50] and is shown as d_2.

The tonal power ratio: The ratio of tonal power is a crucial property for analyzing voice

signals since it determines the signal's spectrum. The link between the power of the signal as a whole and the power of its tonal components is measured by this ratio. The feature is represented by the size [1 × 1] and is denoted as d_3.

Radial flow: A window's spectral flux indicates the degree to which its neighbors' spectra are correlated with one another. The symbol for it is d_4, and its dimensions are [1 × 1].

Weakness in the spectrum: This characteristic measures how skewed the distribution of probabilities is with respect to the mean. Its size is [1 × 1], and it is represented as d_5.

The spectral centroid is a measure of the brightness of sound waves and an indication of where the spectrum is concentrated. This characteristic, with dimensions of [1 × 1], is represented as d_6. The Deep RNN classifier's feature vector is defined as:

$$d = \{d_1, d_2, \ldots . d_6\} \qquad (3.22)$$

With dimensions of [1 × 10^1], the feature vector containing six separate characteristics retrieved from EEG and physiological data is represented by d in this context. To aid in the process of emotion recognition classification, the Deep RNN classifier takes this feature vector, d, as input.

3.14.3 Deep RNN-based Facial Emotion Recognition

The Deep RNN classifier takes the input is a feature vector y that is obtained from physiological signals and EEG. An advancement on the classic RNN, a deep recurrent neural network (DRNN) processes sequential input. To capture complicated patterns and relationships inside sequences, it uses numerous hidden layers piled on top of each other. Natural language processing and voice recognition are two areas that greatly benefit from this design because of the critical need of modeling long-range relationships. While alternative architectures, like as Transformers, have been popular for sequential data workloads, training deep RNNs may be difficult because of problems like vanishing gradients.

- **Continued Relationships:** Recurrent connections within the buried layers of a Deep RNN are its key property. Because of these recurrent connections, the network is able to remember and transfer data from earlier time steps or states to the present one. Important for sequential data because it lets the model account for input sequences' temporal relationships.

- **Dealing with Different Input Lengths:** The capacity to process inputs of varied durations is a major strength of RNNs, particularly Deep RNNs. Random neural networks (RNNs) are great for signal processing, time series analysis, and other applications that don't need input vectors of a fixed size, unlike classic feed-forward neural networks. The recurrent connections are responsible for this adaptability; they may change in length depending on the input sequence.

- **Contextual Understanding:** To produce results in the next state during classification, a Deep RNN takes as input the knowledge of the prior state, which is the previous time step's concealed state. This makes it possible to model to take into account the background and data from earlier time steps when producing forecasts. This comes in handy when looking at sequential data, since the significance of the current data point could be influenced by previous observations.

- **Proven Feature Processing:** Deep RNNs have made a name for themselves by efficiently handling features derived from sequential input, such as signals. Deep RNNs are great at catching patterns and relationships in sequential data, which is essential for signal analysis, prediction, and classification, since they remember past states and use this knowledge in future ones.

The length of the sequential input data might vary; it is received by the input layer. One step at a time, the elements of the sequence are processed.

Deep RNNs have recurrent units or cells in their buried layers. An output and an internal, concealed state are the two primary parts of any unit. Data is carried over via the hidden state from one time step to the next. By stacking these hidden layers, Deep Recurrent architectures can be built, enabling the network to grasp intricate data dependencies.

After all the information has been processed by the hidden layers, the output layer is responsible for producing the final prediction or classification result. Task specificity determines the number of output units; examples of such tasks are regression, multi-class classification, and binary classification.

As shown before, the recurrent connections between hidden units are the most important part of a Deep RNN. In order for the model to keep track of context and temporal patterns,

these linkages make sure that information flows from one phase to the next. In order to train a deep RNN, one may use gradient descent or backpropagation through time (BPTT). In order for the network to produce reliable predictions on sequential data, it is necessary to optimize its weights such that a loss function is minimized.

Deep RNNs are often excellent at processing signals, interpreting natural languages, recognizing voices, and other tasks that involve sequential data because to their structure and properties. Their strength lies in their ability to capture and use temporal dependencies, which greatly enhances their utility in deep learning and machine learning applications.

A network design with several hidden layers that use recurrent connections is the Deep RNN. The classifier can handle inputs of different lengths regarding information sequences because to its architecture. Hence, throughout the categorization process, outcomes in the succeeding state are generated by inputting past state information. The classifier is able to make good use of the characteristics that are given to it because of its recurrent structure. Among the many signal-derived characteristics, Deep RNN stands out as an exceptionally efficient processor. As seen in Figure 3.11, Deep RNN is structured.

At time step p-th, the b-th layer's input is shown as a series of values, allows one to see the structure of Deep RNN, $A^{(b,p)} = \{A_1^{(b,p)}, A_2^{(b,p)}, \dots A_q^{(b,p)} \dots\dots A_L^{(b,p)}\}$ in addition to the data produced by the same layer at step p-th: $B^{(b,p)} = \{B_1^{(b,p)}, B_2^{(b,p)}, \dots B_q^{(b,p)} \dots\dots B_L^{(b,p)}\}$ Each input and output pair is called a unit in this context. On top of that, q is for an arbitrary count of units and s is for the number of units in the ninth tier. In addition, the (b-1) th layer's arbitrary unit count is represented by qand L stands for the quantity of units in that stratum. Accordingly, the weight responsible for transmitting data from the (b-1) th layer to the bth layer is shown as $K^{(b)} \epsilon\, D^{s \times L}$. A collection of weights is represented by L here. In addition, the different parts of the input vector are defined as: G_2^2

$$A_q^{(b,p)} = \sum \; f_{qb}^{(b)} S_b^{(b-1,p)} + \sum \; g_{qq}^{(b)} S_q^{(b-p-1)} \qquad (3.23)$$

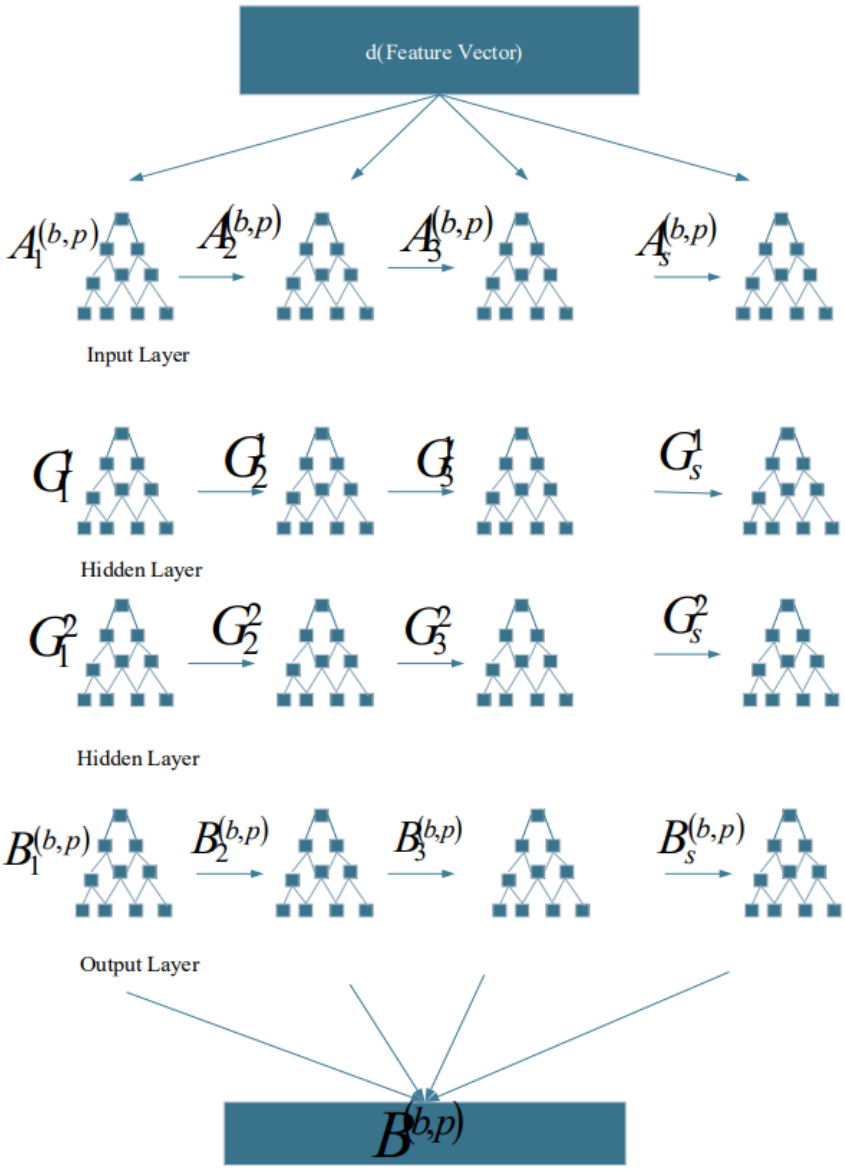


Figure 3.11: Structure of Deep RNN

3.14.3.1 Proposed Algorithm for Optimizing Bat-Rider

The Butterfly–Rider Optimization Algorithm (BROA) is employed to enhance feature selection and classifier parameter tuning in the proposed multimodal emotion recognition framework. BROA is inspired by the foraging behavior of butterflies and rider-based search agents, enabling an effective balance between exploration and exploitation.

Let the population of candidate solutions be represented as:

$$X=\{x_1, x_2,\ldots,x_n\} \qquad (3.24)$$

where each x_i denotes a potential feature subset and model parameter configuration.

The fragrance intensity f_i guiding the butterfly's movement is defined as:

$$f_i = c \cdot I_i^a \tag{3.25}$$

Where

c is the sensory modality constant,

I_i is the fitness value of the solution, and

a is the power exponent determining fragrance absorption.

The global position update is given by:

$$x_i^{t+1} = x_i^t + r \cdot (g^* - x_i^t) \cdot f_i \tag{3.26}$$

Where

g^* is the global best solution,

r ∈ [0,1] is a random number.

For local search (Rider Optimization component), the update rule is:

$$x_i^{t+1} = x_i^t + \alpha \cdot (x_j^t - x_k^t) \tag{3.27}$$

Where

x_j and x_k are randomly selected neighboring solutions and α is the rider acceleration coefficient.

The fitness function F is defined as:

$$F = \max(Accuracy + Precision + Recall) \tag{3.28}$$

The optimization continues until convergence criteria or maximum iterations are reached.

The suggested optimization method BROA is used to train DBN and Deep RNN are two deep learning classifiers. The acronym BROA stands for "ROCA" and "BA" when combined. By mimicking the echolocation behavior of bats, BA is able to detect barriers and possible food sources at great distances. Bats hunt for food by altering the emission rates and pulse wavelengths as they flit through the air at a predetermined speed and frequency. Here, four types of riders are involved in ROA: those who bypass, those who follow, those who assault, and those who overtake. In order to reach the finish line first in a race, each team has a unique strategy. At the end of each time step, the leading rider is chosen according to their success rate. To solve optimization issues, ROA uses the notion of fictional computing, which involves combining ideas and thoughts from the user's imagination. Classification accuracy is improved by combining the parametric features of the two optimization methods. The following is a full outline of the processes involved in the proposed BROA approach:

The first thing to do is to create four groups of riders, let's call them G, and then randomly assign their locations as seen below.

$$H_x = \{H_x(n,.l)\}; 1 \le n \le \phi; 1 \le l \le \psi \qquad (3.29)$$

The number of riders, represented by the symbol ϋ in this case, is the same as the group written as G. The count of coordinates is represented by the symbol $H_x(n,l)$ is equal to n. During the time period x, l indicates where the nth rider was. A rider's SA indicates steering angle, GR indicates gear, AL indicates accelerator, and BK indicates brake.

Finding the best answer to the optimization problem is done by using the fitness function. Here is how the fitness measurement is expressed: The answer with the least amount of mistake is the best one value:

$$E = \frac{1}{N}\sum_{\sigma=1}^{N} [B_\sigma - C_\sigma]^2 \qquad (3.30)$$

Here, N stands for the overall quantity of samples, B_σ signifies the intended result, while C_σ is the deep learning classifier's output, with C being a member of the set {M, M}.

- **Revised answer for riders**

It is common practice for riders to shift their bodies in order to better see the leader of a group and follow their cues. Some of these characteristics include the rider's prominent position in the group, distinctive attire or gear, nonverbal clues, riding technique, and communication signals. In team sports like cycling or horseback riding, knowing who is in the lead is essential for communication, planning, and safety.

- **Revise the calculation for the bypass rider**

In most cases, the bypass rider will not follow the leading rider's trajectory and will instead steer clear of it. The following equation, which updates the position, reflects this behavior:

$$H_{x+1}^{ay}(n,l) = \gamma\left[H_x(\tau,l) * g_1(l) + H_x(g_2,l) * [1 - g_1(l)]\right] \quad (3.31)$$

In this case, y and g_i stand for random integers between [0,1], while τ and g_2 stand for random integers between 1 and ö, respectively.

- **Bring the follower equation up to date:**

To approach the destination, the follower rides in a manner similar to a horse, adjusting its position in response to the leader's movements. In addition, the following is the expression for the position update equation:

$$H_{x+1}^{hl}(p,q) = H^{Mn}(MN,q) + \lfloor Cos\left(SA_{p,q}^{*}\right) * H^{MN}(MN,q) * S_p^x \rfloor \quad (3.32)$$

$$H_{x+1}^{hl}(p,q) = H^{Mn}(MN,q)\left(1 + Cos\left(SA_{p,q}^{*}\right) * S_p^x\right) \quad (3.33)$$

Where $H^{*} = H^{Mn}(MN,q)$

Understanding emotions via different mediums, such as written words, spoken word, and facial expressions, relies heavily on BA. The purpose of BA is to integrate and assess data from many sources in order to increase the effectiveness of deep learning models for emotion identification. Systems for recognizing emotions, which might be helpful in fields like mental health assessment and human-computer interaction, can be greatly enhanced by this data fusion.

$$H_{x+1}^{hl}(p,q) = H_x^{hl}(p,q) + P_x(p,q) \quad (3.34)$$

$$H_{x+1}^{hl}(p,q) = H_x^{hl}(p,q) + P_{x+1}(p,q) - (H_x^{hl}(p,q) - H^{*})G(p,q) \quad (3.35)$$

$$H^{*}G(p,q) = H_{x+1}^{hl}(p,q) - H_x^{hl}(p,q) - P_{x+1}(p,q) + H_x^{hl}(p,q)G(p,q) \quad (3.36)$$

$$H^{*}\frac{H_{x+1}^{hl}(p,q)-H_{x}^{hl}(p,q)-P_{x+1}(p,q)+H_{x}^{hl}(p,q)G(p,q)}{G(p,q)} \quad (3.37)$$

The derived H^{*} is used.

$$H_{x+1}^{hl}(p,q)=\frac{H_{x+1}^{hl}(p,q)-H_{x}^{hl}(p,q)-P_{x+1}(p,q)+H_{x}^{hl}(p,q)G(p,q)}{G(p,q)}(1+Cos(SA_{p,q}^{*}) * S_{p}^{*} \quad (3.38)$$

$$H_{x+1}^{hl}(p,q)=\frac{H_{x+1}^{hl}(p,q)}{G(p,q)}(1+Cos(SA_{p,q}^{*}) * S_{p}^{*} - \frac{H_{x+1}^{hl}(p,q)+P_{x+1}(p,q)-H_{x}^{hl}(p,q)G(p,q)}{G(p,q)}(1+Cos(SA_{p,q}^{*}) * S_{p}^{*}) \quad (3.39)$$

$$H_{x+1}^{hl}(p,q)-\frac{H_{x}^{hl}(p,q)}{G(p,q)}(1+Cos(SA_{p,q}^{*}) * S_{p}^{*} = -\frac{H_{x}^{hl}(p,q)G(p,q)-P_{x+1}(p,q)-H_{x}^{hl}(p,q)}{G(p,q)}(1+Cos(SA_{p,q}^{*}) * S_{p}^{*}) \quad (3.40)$$

$$H_{x+1}^{hl}(p,q)\left(1-\frac{(1+Cos(SA_{p,q}^{*}) * S_{p}^{*})}{G(p,q)}\right) = \frac{H_{x}^{hl}(p,q)G(p,q)-P_{x+1}(p,q)-H_{x}^{hl}(p,q)}{G(p,q)}(1+Cos(SA_{p,q}^{*}) * S_{p}^{*} \quad (3.41)$$

$$H_{x+1}^{hl}(p,q)\left(\frac{G(p,q)-1-Cos(SA_{p,q}^{*}) * S_{p}^{*}}{G(p,q)}\right) = \frac{H_{x}^{hl}(p,q)G(p,q)-P_{x+1}(p,q)-H_{x}^{hl}(p,q)}{G(p,q)}(1+Cos(SA_{p,q}^{*}) * S_{p}^{*} \quad (3.42)$$

$$H_{x+1}^{hl}(p,q)\frac{G(p,q)}{(p,q)-1-Cos(SA_{p,q}^{*}) * S_{p}^{*}}\left[\frac{H_{x}^{hl}(p,q)G(p,q)-P_{x+1}(p,q)-H_{x}^{hl}(p,q)}{G(p,q)}(1+Cos(SA_{p,q}^{*}) * S_{p}^{*}\right] \quad (3.43)$$

Where

$$G(p,q)=G(Gmin_{max}N_{3})_{min} \quad (3.44)$$

G = p * q This stands in for the G function, which takes two parameters, P and q. G (Gmin$_{max}$ N$_3$) Another function is involved in this portion of the equation, Gmin$_{max}$ N$_3$, it is used on the input. The rider's position index.

$SA_{p,q}^{x}$ Visible at P and q is the steering angle.

S_{p}^{x} Distance traveled from point P to the riders.

- **Update the over taker solution**

The over taker's position update equation takes into account several variables that impact its trajectory. Among these considerations is the over taker's fitness level, which is indicative of its potential for navigating or adapting to its existing surroundings. The over taker may change its location in a certain dimension or direction (e.g., up-down, left-right) by using the coordinate selector. Depending on whether the over taker should go forward or back along the chosen coordinate, the direction indication shows them how to proceed. All of these things work together to provide the over taker a method for constantly adjusting its position according to its fitness and directions, which is crucial for maximizing its movement strategy in any situation, like a race or a navigational challenge.

$$H_{x+1}^{ot}(p,q) = H_x(p,q) + [I_x(p) \times H^{MN}(MN,q)] \qquad (3.45)$$

The passing is done as $H_{x+1}^{ot}(p,q)$. This computation incorporates indications depending on the direction that is taken into account as I_x (p) throughout the area P.

- **Equation updating Attacker**

In a competitive setting, the attacker's position or plan may be updated using an equation that represents their interactions with their surroundings and rivals mathematically. Typical components include the following: the attacker's present location; the leader's position; the attacker's speed; direction; and strategy. To mimic how the attacker's location changes over time, time steps are often used. In order to simulate the intricacies of the actual world, it is possible to include environmental elements and control inputs. In order to approach or exceed the leader, this equation is a great tool for optimizing the attacker's moves and forecasting competitive strategies and tactics in many circumstances. Equation used by the assailant $H_x^{att}(p,n)$ appears below.

$$H_x^{att}(p,n) = H^{MN}(MN,n) + \left[cos\ cos\left(SA_{p,q}^{x}\right) \times H^{MN}(MN,n) * S_p^x\right] \qquad (3.46)$$

- **Feasibility Calculation**

To find out whether a project can be done, you must first assess its viability. Knowing how successful it may be requires looking at a variety of factors, including technical, financial, operational, legal, scheduling, and market considerations. Through the

identification of risks and opportunities and the verification that the project is in accordance with all applicable legislation and resources, this analysis aids stakeholders in making well-informed choices.

- **Finalizing**

Until the intended outcome is achieved, this process iteratively keeps going. The process entails iteratively working toward the goal by doing activities, making modifications, and evaluating progress. To get the intended result, methods and efforts may be fine-tuned using this method, which permits adaptation and refining. It is an adaptable and dynamic approach that seeks to increase the probability of reaching the desired goal.

3.15 ALGORITHM FOR MULTIMODAL EMOTION RECOGNITION USING DEEP LEARNING NETWORKS

First things first in the suggested BROA-based deep learning algorithm: initializing parameters including learning rates, population size, and generations. It keeps an eye on fitness and convergence while it generates a population of neural networks via mutation, selection, and crossover operations. After the final population is determined, the network is trained using the dataset with the greatest performance. Predictions about fresh data are then generated by this trained network.

The method.1 Fusion Techniques Based on Deep Learning Networks for Multimodal Emotion Recognition

Start

{

Input video phase ()

{

int H, A, n;

//Import the initiation parameters and data into the dataset.

$H = \{A_1, A_2 \ldots A_i \ldots\ldots A_n\}$; *//using eqn.(1)*

//Import database

}

Feature analysis module ()

{

int $L_m, c_r, d_r, \lambda, l_c, l_t$

;

//features from the pre-processed images were considered.

$L_m(c_r, d_r) = \sum_{n=1}^{7} \lambda (l_c - l_t) 2^{\beta n}$ *//using eqn.(3)*

//meaningful features were extracted

}

Classification module ()

{

int X, n ;

//using DBN module

$X^1 = \{X_1^1, X_2^1, \ldots\ldots\ldots X_i^1 \ldots\ldots X_n^1\}$; *// using equation (10)*

//Emotions were taken

Finding Emotions with signals ()

{

int

// using deep RNN

$X_q^{(b,p)} = \sum f_{qb}^{(b)} S_b^{(b-1,p)} + \sum g_{qq}^{(b)} S_q^{(b,p-1)}$;

//predict the required

}

Training ()

{

int

// training using bat rider

$$H_{x+1}^{hl}(p,q) = \frac{G(p,q)}{G(p,q) - 1 - Cos\left(SA_{p,q}^{x}\right) * S_p^x} \left[\frac{H_x^{hl}(p,q) G(p,q) - P_{x+1}(p,q) - H_x^{hl}(p,q)}{G(p,q)} \left(1 + Cos\left(SA_{p,q}^{x}\right) * S_p^x\right) \right]$$ *//using equation* (39)

// the required is obtained

```
}
Synthesize ()
{
    Int
    //using Renyi Fusion Entropy
    R = 1/(1-d) log(E)                    //using equation (43)
    //required emotions were obtained
}
}
Stop
```

Figure 3.12: Proposed Algorithm

3.16 FLOW CHART OF PROPOSED BROA-DEEP NN

It is a methodical procedure to use deep learning networks and fusion techniques for multimodal emotion recognition. It starts by taking in many forms of data, such as text, video, and audio, all of which stand for different aspects of expression of emotion. Deep recurrent neural networks (DBNs), Transformers, and Deep RNNs are some of the specialized deep learning models that handle these input streams independently. The fusion step then uses either early or late fusion procedures to combine their outputs. In the end, the combined data is utilized for accurate Emotion Recognition, which uses the insights from all modalities to provide a reliable emotional evaluation. The efficiency and reliability of emotion detection systems are both improved by this simplified method.

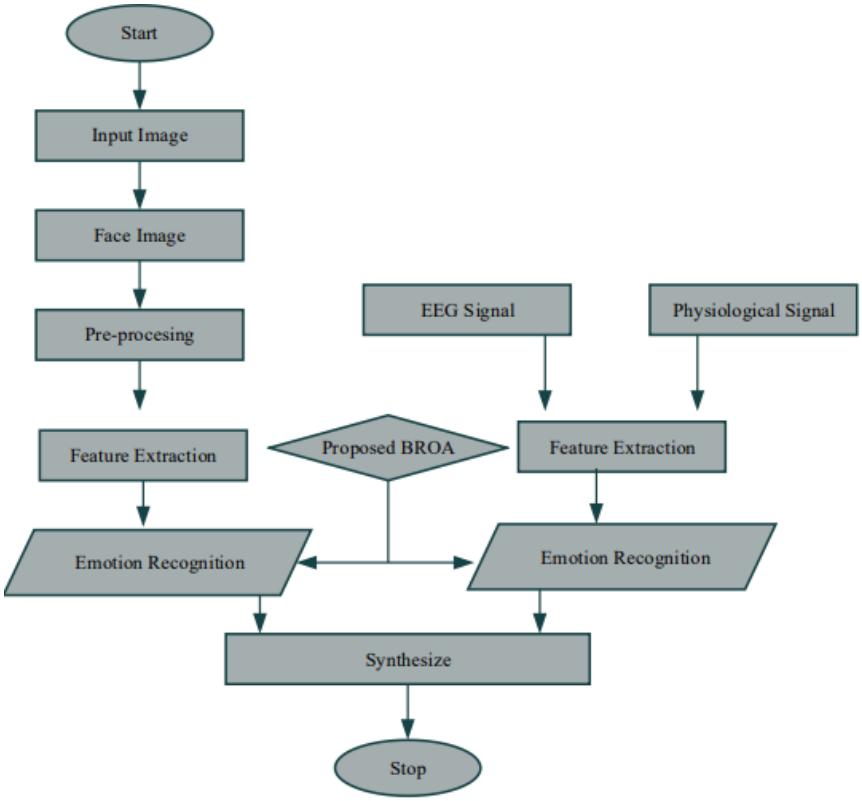


Figure 3.13: Flow Chart of the Proposed BROA-Deep NN

The BROA-DNN was adopted as the basic deep learning framework in this study. Practically, the architecture was incorporated into the experimental pipeline by initially implementing the preprocessing steps that have been discussed above (resizing inputs, normalization, and augmentation). The data underwent processing and then it was ingested into the BROA-DNN where the layered nature of the model model made features learn in hierarchical order- starting with low-level edge and texture representation, as well as, high-level semantic representations of expressions.

In order to fit the BROA-DNN to our experiments, the last fully connected layers were adjusted to fit the target classification categories used in Chapter 3. Transfer learning was used where it was feasible, in which case the pretrained feature extractors were fixed during the classification head training. This made the BROA DNN able to retain general representations and be specialized in the finer details of emotion recognition.

In experiments (described in Chapter 4), BROA DNN was trained and fine-tuned on the curated dataset of emotional samples. The evaluation of performance was conducted under varying conditions of learning rates, optimizers and dropout rates to determine the best

training strategy. The application of the BROA DNN to these experiments indicated that it is able to extrapolate and classify a wide expression, which helps to justify its applicability when it comes to the emotion detection problem.

Once the BROA DNN architecture has been defined, the model has been implemented into an experimental setup systematically. As the input, the preprocessed dataset (resized to 48×48 pixels, normalized, and augmented) was used in the first place. These images were fed through the BROA DNN model with the convolutional layers having hierarchical features of facial expression. To fit the model to our study, the BROA-DNN is fully connected layers that were customized to fit our study, i.e., the emotion categories.

We have used this modified structure as we trained BROA-DNN in our experiments where parameters like learning rate, batch size and dropout were adjusted to ensure stability and accuracy. Several runs were made to compare performance in various configurations and they were noted down in the form of classification accuracy and loss value. This experimental implementation showed how the BROA-DNN had not just been conceptually implemented, but implemented as well, to test and categorize emotional data into our study.

3.17 SYSTEM IMPLEMENTATION

With its ability to independently extract meaningful features from pictures via a sequence of convolutional and pooling procedures, the VGG-16 CNN, as seen in Figure 3.14, has been trained on a large and diverse picture dataset. Using VGG-16, this model has ImageNet weights that have already been trained. More categorization layers will be added to the suggested technique. The next step, after creating a one-of-a-kind model, is to test and improve it using a designated dataset. What follows is a discussion of the CNN layers.

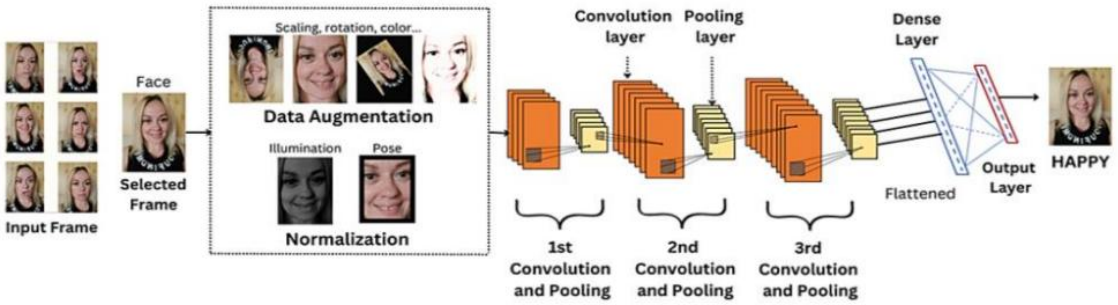


Figure 3.14: Emotion recognition using VGG16-CNN

3.18 OVERALL ARCHITECTURE OF THE PROPOSED SYSTEM

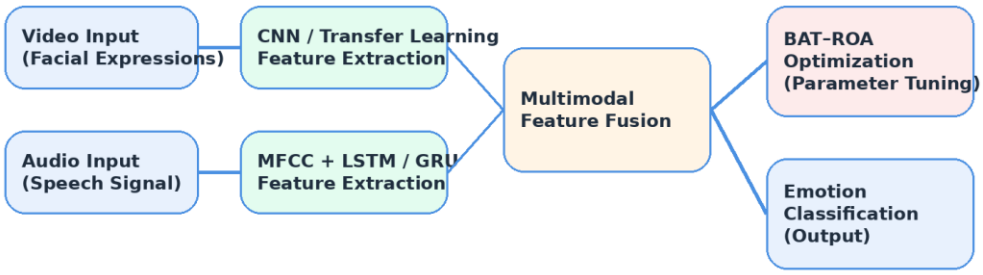


Figure 3.15: The complete architecture of the proposed multimodal emotion recognition system is illustrated, highlighting the video-based CNN module, audio-based LSTM/GRU module, feature-level fusion, and hybrid BAT–ROA optimisation.

The proposed multimodal emotion recognition system consists of three primary modules: the Video Emotion Recognition Module, the Audio Emotion Recognition Module, and the Multimodal Fusion and Optimization Module. The overall architecture is illustrated in Figure 3.15.

Video Emotion Recognition Module

The video module processes facial images extracted from video frames. Each input facial image is resized to a dimension of 48 × 48 × 1 and fed into a Convolutional Neural Network (CNN). The CNN architecture consists of multiple convolutional and pooling layers that progressively extract spatial features.

- **Input Layer:** 48 × 48 × 1
- **Convolution Layer 1:** 32 filters, 3 × 3 kernel → Output: 48 × 48 × 32
- **Max Pooling Layer 1:** 2 × 2 → Output: 24 × 24 × 32
- **Convolution Layer 2:** 64 filters, 3 × 3 kernel → Output: 24 × 24 × 64
- **Max Pooling Layer 2:** 2 × 2 → Output: 12 × 12 × 64
- **Flatten Layer:** 9216 neurons

- **Fully Connected Layer:** 128 neurons

The output of the video module is a 128-dimensional feature vector representing facial emotion features.

Audio Emotion Recognition Module

The audio module processes speech signals using Mel-Frequency Cepstral Coefficients (MFCCs). Speech signals are segmented and transformed into MFCC feature vectors before being fed into a recurrent neural network.

- **Input Audio Signal:** Speech waveform
- **MFCC Extraction:** 40 MFCC coefficients per frame
- **Input Feature Dimension:** T × 40 (T = number of frames)
- **LSTM / GRU Layer:** 128 hidden units
- **Dense Layer:** 128 neurons

The audio module produces a 128-dimensional temporal feature representation of speech emotions.

Multimodal Fusion and Optimization Module

The feature vectors obtained from the video and audio modules are concatenated to form a unified multimodal feature representation.

- **Feature-Level Fusion:**

$$F_\{"\{fusion\}\} \; = \; [F_\{"\{video\}\} \; ; \; || \; ; \; F_\{"\{audio\}\}]$$

- **Fusion Feature Dimension:** 256

The fused feature vector is passed to the Hybrid Bat–Rider Optimization Algorithm (BROA), which optimizes fusion weights and classifier parameters. The optimized feature representation is then fed into a fully connected classification layer.

- **Fully Connected Layer:** 64 neurons
- **Output Layer:** Softmax classifier with *N* emotion classes

3.18 HARDWARE AND SOFTWARE ENVIRONMENT

The experimental evaluation of the proposed multimodal emotion recognition system was carried out on a standard computing platform with GPU support to ensure efficient training and evaluation of deep learning models. The hardware and software specifications used for implementation are summarized in Table 3.7.

Table 3.7: Hardware and Software Configuration

Component	Specification
Processor	Intel Core i7 / Equivalent
RAM	16 GB
Storage	1 TB HDD / 512 GB SSD
GPU	NVIDIA GPU with CUDA support
Operating System	Windows 10 / Ubuntu Linux
Programming Language	Python 3.8
Deep Learning Framework	TensorFlow 2.x with Keras
Audio Processing Library	LibROSA
Image Processing Library	OpenCV
Optimization Algorithm	Hybrid BAT–ROA (Custom)
Development Environment	Anaconda / Jupyter Notebook

3.19 TOOLS AND FRAMEWORKS USED FOR IMPLEMENTATION

The proposed multimodal emotion recognition system and hybrid BAT–ROA optimization framework were implemented using widely adopted programming languages, deep learning libraries, and development tools. These tools were selected to ensure computational efficiency, scalability, and reproducibility of experimental results.

Table 3.8: Software Tools and Frameworks Used

Tool / Framework	Version	Purpose
Python	3.8 / 3.9	Core programming language for model development
TensorFlow	2.x	Deep learning framework for CNN and LSTM models
Keras	Integrated with TensorFlow	High-level API for neural network design
OpenCV	4.x	Image preprocessing and

		face detection
NumPy	1.x	Numerical computation and array operations
Pandas	1.x	Data handling and preprocessing
LibROSA	0.9.x	Audio signal processing and MFCC extraction
Scikit-learn	1.x	Performance evaluation and utility functions
Matplotlib	3.x	Visualization of results and performance graphs
Jupyter Notebook	Latest	Experimental prototyping and analysis

Table 3.9: Development Environment and Framework Support

Component	Description
Operating System	Windows 10 / Ubuntu Linux
IDE / Environment	Anaconda / Jupyter Notebook
Deep Learning Backend	TensorFlow (GPU-supported)
Audio Processing	LibROSA-based MFCC extraction
Optimization Module	Custom implementation of BAT–ROA
Visualization	Matplotlib-based plots

3.19 DATASET DESCRIPTION AND DISTRIBUTION

The proposed multimodal emotion recognition system was evaluated using publicly available benchmark datasets corresponding to different modalities. Facial emotion data, speech emotion data, and physiological signals were sourced from standard datasets to ensure experimental validity and reproducibility. The distribution of datasets used in this research is summarized in Table 3.10.

Table 3.10: Dataset Distribution and Usage

Dataset Name	Modality	Number of Samples	Emotion Classes	Purpose
FER-2013	Facial Images	~35,000 images	7 (Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral)	Training & Testing Video Emotion Model
RAVDESS	Speech Audio	1,440 audio files	8 (Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, Surprise)	Training & Testing Audio Emotion Model
DEAP	Physiological Signals	32 subjects × 40 trials	Valence–Arousal based	Multimodal Validation &

- **FER-2013** was used exclusively for training and evaluating the CNN-based facial emotion recognition module.
- **RAVDESS** speech recordings were utilized for extracting MFCC features and training the LSTM/GRU-based audio emotion recognition module.
- **DEAP** dataset was used to validate multimodal emotion recognition performance and to support physiological emotion analysis.

All datasets used in this research are publicly available, and no primary data collection involving human subjects was conducted.

CHAPTER – 4
EXPERIMENTAL RESULTS AND DISCUSSIONS

4.1 OVERVIEW

This chapter explores how the training was put into practice and application modules, with the training module serving primarily to train the DCNNs. The second part of the chapter details the recommended system modules. Lastly, the other component includes the decisions taken and experiments conducted during training to determine the ideal hyper parameters and DCNN model, as well as experiments conducted during test to evaluate the models' accuracy.

4.2 DATASET OUTLOOK

Many different deep learning and The FER2013 dataset has been used to train and assess machine learning models for facial emotion identification. In this research, the FER 2013 dataset was preprocessed by generating a CSV file. This file included the following attachments: labels for values and feelings for picture pixels, and a use column that distinguished between the training and validation sets of images. The collection is comprised of 35,887 grayscale pictures with dimensions of 48x48 pixels. These photographs are categorized into seven groups, each representing a particular emotion. Our facial expression analysis models are trained and evaluated on this dataset often during this investigation. This is how the resume appears:

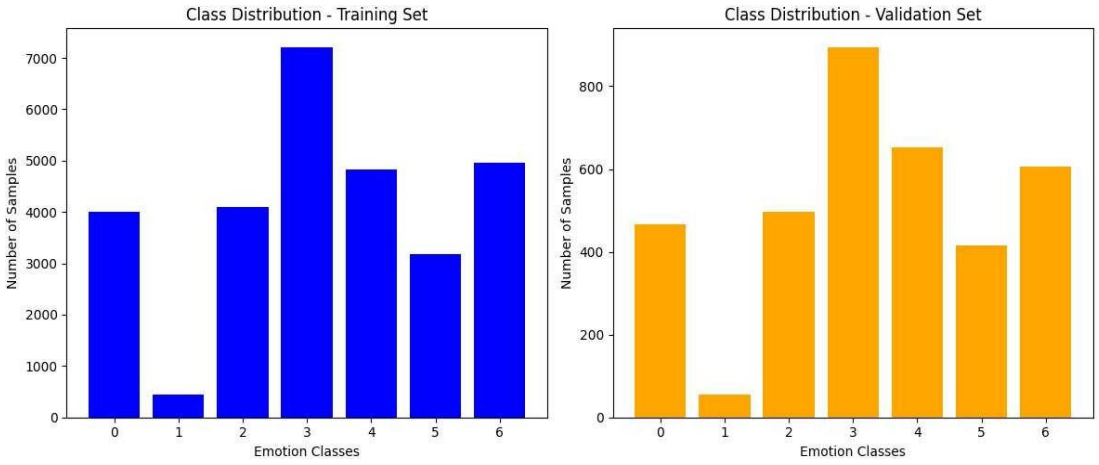


Figure 4.1: Class distribution for the validation and training sets, respectively.

In the columns of the Fer2013 dataset, you may find label mappings with seven different emotions. This mapping is utilized: A score of Anger is represented by 0; disgust by 1; fear by 2; pleasure by 3; grief by 4; surprise by 5; and neutrality by 6. This research considered validation and utilized 22968 for training.

4.3 EXPERIMENTAL SETTING

.

Before I could build and train an emotion detection Preprocessing the dataset was necessary for the Convolutional Neural Network (CNN) model by making sure all the photos were the same size, which was 48 by 48 pixels which is given as follows in the pictures.

The dataset utilized in this research included facial expression samples that consisted of Neutral, Happy, Surprised, and Contempt, as presented in the exemplary figures (Figures 4.2-4.13). These facial expression samples were utilized to develop the emotion detection model. In order to process the dataset, a structured architecture was needed, so the following section (Architecture of Emotion Detection Models) describes how the images were preprocessed in a standard image size of 48x48 pixels, and the modifications to use transfer learning to train a model with the dataset. The pretrained VGG16 model was set as the base architecture in which the convolutional layers were used as a feature extractor and a new classification head designed specifically for this emotion model used for classifying the target emotion. Accordingly, the experimental samples informed the architecture of the Emotion Detection Model and the training method discussed in subsequent sections.

To train a transfer learning model to classify emotions in faces, I first loaded the pre-trained models for Google Net(V3), ResNet50, and VGG-16, and then froze all of their layers. A new, fully connected layer was created on VGG-16 with 48 units, and a SoftMax layer for classification was added. Furthermore, the layers of the Hall model pretrained must be "frozen," or have their weights kept the same during training. The reasoning for freezing

the pretrained layers is twofold: it speeds up training and it prevents overfitting to the training set.

Figure 4.2: Neutral





Figure 4.4: Surprised



Figure 4.3: Happy



Figure 4.5: Happy



Figure 4.6: Student Happy



Figure 4.7: Student Happy



Figure 4.8: Student Contempt



Figure 4.9: Student Happy



Figure 4.10: Student Surprise



Figure 4.11: Student Neutral



Figure 4.12: Student Neutral



Figure 4.13: Student Happy

Next, I trained the model with the same validation and training sets as the DCNN models. Again, the layers of the model may stay frozen during training while the weights of the SoftMax layer and the newly connected layer are changed. This technique trains the model to obtain new representations that are specific to emotion recognition, but using the trained weights from ImageNet. After training, the three transfer learning model performances of ResNet50, VGG16, and Google Net (V3) can be collected and compared using the data from the validation set.

4.3.1 Preprocessing the dataset

The dataset should be preprocessed before developing and training an emotion detection model. CNNs take structured and uniform inputs; therefore, all images were scaled to 48 × 48 pixels. All samples which were put into the CNN maintained the same spatial dimensions by resizing all the images to the same size. This assisted in eliminating variation in sampling of various images of varying resolutions. The preprocessing would not have enabled the model to learn useful features. The reason is that the various spatial sizes would have led to feature maps which could not be utilized and the amount of computation would have been unpredictable.

Besides resizing, preprocessing aids in the minimization of the computational complexity. The reduction of smaller picture sizes, which are 48 x 48 pixels, reduces the parameters that the CNN has to deal with considerably. This enables the faster training yet retaining adequate amount of facial feature (eyes, eyebrows, mouth and movement of facial muscles etc.) which is essential in determination of emotions like joy, surprise, disgust or indifference.

Every emotion recognition model uses a different architecture specific to the aspect of face expression recognition. This section will give a specific discussion of the most popular pre-trained model architectures and how they can be adapted to emotion detection tasks.

4.3.1.1 Pipeline Data Preprocessing.

Prior to the construction and training of models to detect emotions, an extensive pre-processing of the dataset is required to achieve the best possible results:

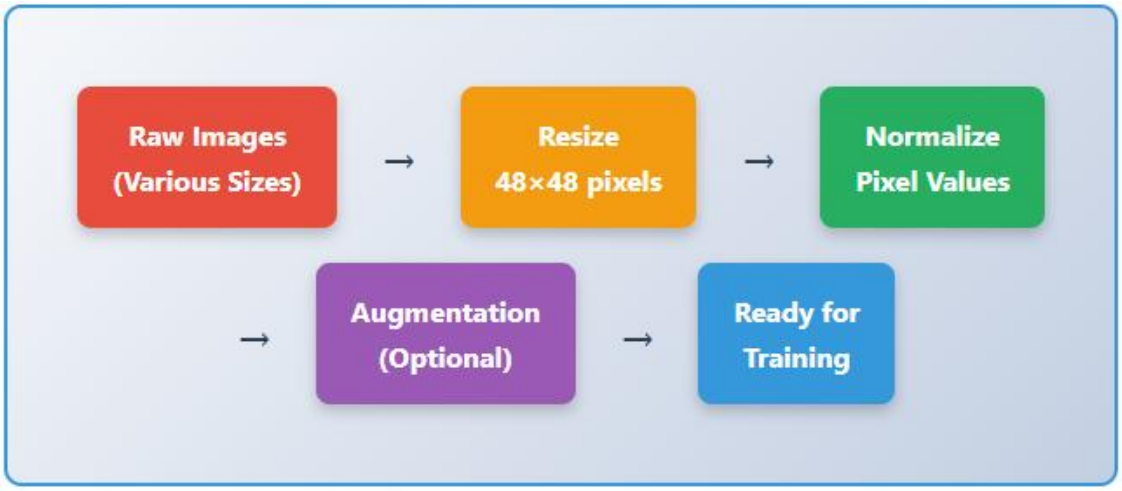


Figure 4.14: CNN Model necessities based on Preprocessing

Steps to be followed:

- ✔ **Image Resizing:** Although we process all the pictures in 48x48 pixels, it is to guarantee consistency in the size of the input picture**.**
- ✔ **Grayscale to RGB Conversion:** Translating Grayscale images to 3 channel RGB to use them through pre-trained models**.**
- ✔ **Pixel Normalization:** Squeeze the pixel values to [0,1] so as to have consistent training.
- ✔ **Data Augmentation:** Rotation, flipping, and changing brightness of the images so that it becomes more diverse.

A pipeline preprocessing preconditions the groundwork on successful training of CNN-based models of emotion detection. The 48×48-pixel processing of all the images would not only ensure consistency across the dataset, but also allow fewer computations to be performed to execute the models, and maintain the meaningfulness of each facial expression of the face. The pretrained models such as VGG16 that take the color images need to be turned into grayscale images and finally turned into 3-channel RGB of images. The gradients become stabilized during the training Pixel normalization (with scaling pixel

values to the range [0,1]) process. Data augmentation (rotating, flipping, and adjusting the brightness) assists in increasing the variety of the dataset. All these enhance the strength of the model and reduce overfitting.

After doing the preprocessing, the second step was to use transfer learning to perform emotion recognition. The amount of time and computational resources needed to train a CNN using just datasets are generally huge. Nevertheless, transfer learning allows learning models to be constructed quickly using information acquired through an already trained model using a large-scale dataset like ImageNet, which consists of millions of images that are different.

The already trained model already represents low-level (edges, corners, textures) and mid-level features (shapes, patterns, outlines in faces). The features are generic and applicable to numerous computer vision applications, such as the detection of emotions. In order to change this network and fit it to this task, an additional classification layer was introduced on top of the trained network. Denoting emotion data, the specified layer was adjusted so that the model would get to learn how to map features extracted to the categories of emotions such as Neutral, Happy, Surprised, or Contempt.

With preprocessing and transfer learning, the model was capable of accomplishing effective training, reduced overfitting and greater accuracy through relatively small amounts of data. This method enabled the CNN to concentrating on minute differences in facial emotions that can differentiate between emotions; it is therefore suitable in the activity of emotion detection.

4.3.1.2 Base Model: VGG16

In this study the VGG16 is picked to use as a base model in transfer learning. VGG16 is a deep Convolutional Neural Network based architecture that was initially trained on the ImageNet dataset that contains more than one million images in 1,000 categories. Its advantage is that it has a simple, but the effective architecture which comprises 16 layers of weight using small $3 \, x \, 3$ convolution filters used in sequence. Such a design allows the model learns hierarchical visual feature where simple edges and textures are learned at the

initial stages to more intricate details in later stages such as shapes and structures of objects.

VGG16 is a useful emotion recognition feature extractor because of its track record of being a strong feature extractor in visual representations that can be transferred to other programs. To reuse the low- and mid-level feature maps (edges, contours, facial regions, etc.), the VGG16 model that has been pretrained was used instead of the network that has to be trained on a significantly larger dataset to be used.

In order to adjust VGG16 to the emotion detection task, the fully connected layers that were intended to be used by ImageNet (classification) were substituted by a new classification head that claimed the emotions types on the dataset (Neutral, Happy, Surprised, Contempt, etc.). This was done in training by keeping the previous convolutional layers frozen to maintain their extracting general features property, and the new classification layers were trained and fine-tuned to extract subtle changes in facial expressions.

In order to adjust the base model to the new area (emotion detection), a new classification layer was installed over the pretrained network. This layer was assigned with a task of classifying the target categories in the dataset, e.g. Neutral, Happy, Surprised, and Contempt. Training the top layers only, the model would learn domain-specific aspects (patterns of facial expressions) rather quickly and maintain the low- and mid-level features acquired in ImageNet.

In that way, the model could be trained on fewer data as opposed to the creation of a CNN in its entirety. It further minimized the time used during training and enhanced accuracy, as the feature extraction skills of the base model were already optimized in such a way. Therefore, the transfer learning using a trained base model offered the effective and strong approach to adjusting to the specifics of facial emotion recognition.

We used VGG16 model that had been trained on ImageNet as the backbone in our experiments. ImageNet features are highly generic and depict low- to mid-level representations (edges, textures, shapes etc.), thus, the latter act as an effective baseline of transfer learning. In order to make VGG16 work in our domain with comparatively very

little data, we kept the convolutional base of the model and inserted a newly fully connection classification layer at its top. This custom classification layer was used to make training with our data and the previous convolutional layers were either frozen or fine-tuned accordingly.

The given method enabled the model to repurpose already learned ImageNet features and learn only the last classification layer which decreased over-fitting as well as computational expense. This means that VGG16 would be able to adapt well to the details of the new domain using a relatively small number of training samples, and at the same time using the highly-representative features that it already learns on ImageNet.

In this manner this held true to the model as it was able to accomplish:

- ✔ The benefits of reduced training time most of the weights were already optimized.
- ✔ Lower data requirement, the pretrained features reduced the demand of huge datasets.
- ✔ Better accuracy, because VGG16 presented powerful feature representations that resulted in a good base in emotion recognition.

Therefore, transfer learning is an effective and efficient method of detecting facial emotions as VGG16 proved to be a strong base model in the given study.

16-layer deep Convolutional Neural Network trained by the Visual Geometry Group at Oxford originally provided on ImageNet with more than 1 million images in 1,000 categories.

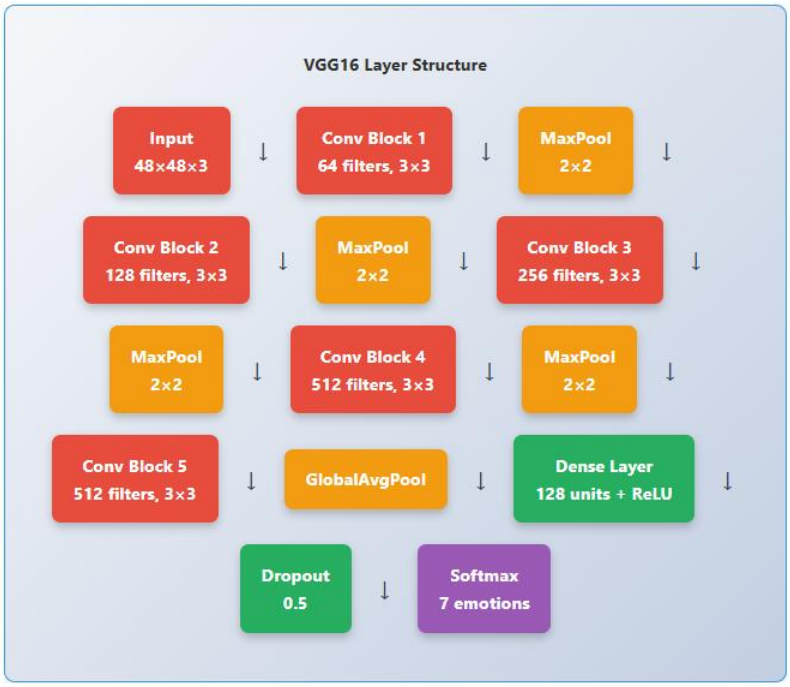


Figure 4.15: VGG16 architecture

The entire VGG16 architecture has the full flow of convolutional and pooling layers, which in turn derive progressively features through the facial image. On the basic layers, low level features like edges, textures are detected and progressively deeper layers capture more shape and face features. VGG16 has high representational capacity through stacked small (3X3) convolutional filters, which are designed with very simple structure. When used in emotion detection, the pretrained layers act as frozen feature detectors, whereas newly added layers are used to encode the emotion into emotion categories.

Table 4.1: Layer Specifications of VGG16

Layer Type	Output Shape	Parameters	Function
Input Layer	48×48×3	0	Preprocessed RGB images
Conv2D (Block 1)	48×48×64	1,792	Edge and texture detection
Conv2D (Block 1)	48×48×64	36,928	Feature refinement
MaxPool2D	24×24×64	0	Spatial dimensionality reduction
Conv2D (Block 2)	24×24×128	73,856	Complex pattern detection
Global Average Pooling	512	0	Feature vector extraction
Dense (Custom)	128	65,664	Emotion-specific learning
Dropout	128	0	Overfitting prevention

Softmax Output	7	903	Emotion classification

This table describes the separate layers of the adapted VGG16 model in the process of recognizing emotions. The input layer is fed with preprocessed images (48x48x3). Blocks 1 and 2 contain convolutional layers that pick up basic visual representations (edges, textures, and complicated patterns). The max pooling layer operates by making the spatial size smaller and important features are retained. A world average pooling layer features maps into a sparse feature list that is after that learned by a special dense layer to acquire emotion-related designs. The dropout is used to reduce overfitting, and ultimately, a softmax output layer was used to classify images to the seven target emotions. This architecture is a balance between knowledge that is pretrained and knowledge that is specific to the task.

To detect emotion in this research:

i. **Input Layer**: Preprocessed grayscale images of the 48×48 size were scaled to 3 channel color (RGB) as per the expected input of VGG16.

ii. **Convolutional Base (Frozen)**: The feature extractor was VGG16 frozen convolutional layers trained using ImageNet dataset. These layers pick the low and mid visual patterns like edges, textures and faces patterns.

iii. **Flatten/Pooling Layer:** The resulting feature maps were subjected to Global Average Pooling layer in order to downsize the dimensionality without losing the important data.

iv. **New Head - Fully Connected Layer:** This is a dense layer consisting of ReLU activation and Dropout to learn emotion specific patterns that will not overfit.

v. **Output Layer**: It included the softmax layer that was utilized to assign the extracted features to the set of emotion categories (Neutral, Happy, Surprised, Contempt, etc.).

The design can be used to give the pretrained VGG16 layers good generic features, and the

new layers used to detect subtle facial expression cues.

4.3.1.3 Implementation of Transfer Learning

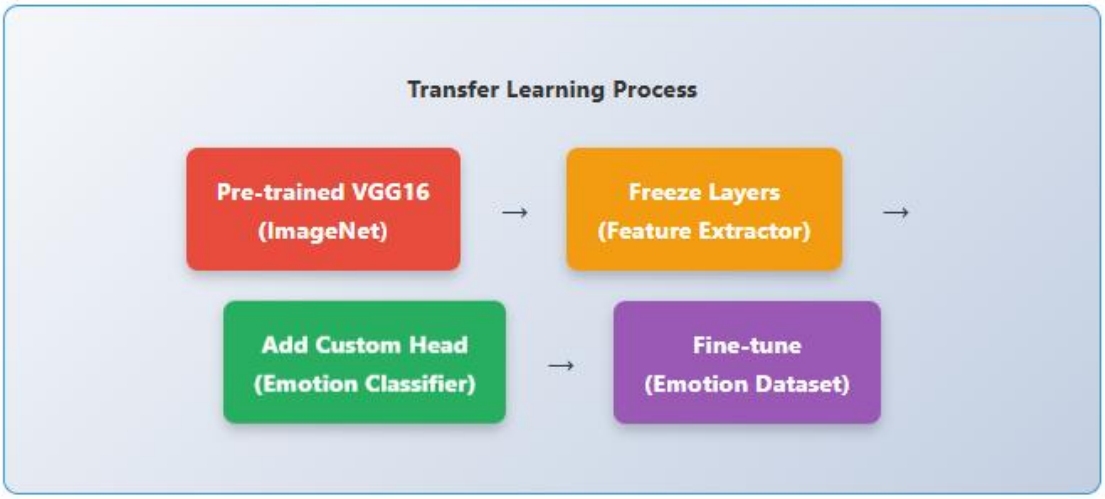


Figure 4.16: Transfer Learning process

The benefits of transfer learning include:

✔ Small Training Time: Pre-trained weights avoid having to train the weights.

✔ Reduced Data Requirements: Capitalize on the available feature representations.

✔ There is better accuracy: Take advantage of the variety of visual patterns of ImageNet.

✔ Like optimized feature extractors: Use more rapid featured Convergence.

The transfer learning schematic lays emphasis on the role of pretrained weight of VGG16 in speeding up the process of learning to detect an emotion. The system has the advantage of strong feature representations because it uses a model pretrained on a massive scale ImageNet dataset. These are the benefits of less time training, less data, and the speed of convergence than training fresh. The training convolutional base captures general image features and the new classification head is trained to identify facial emotions.

4.3.1.4 Comparative Architecture Analysis.

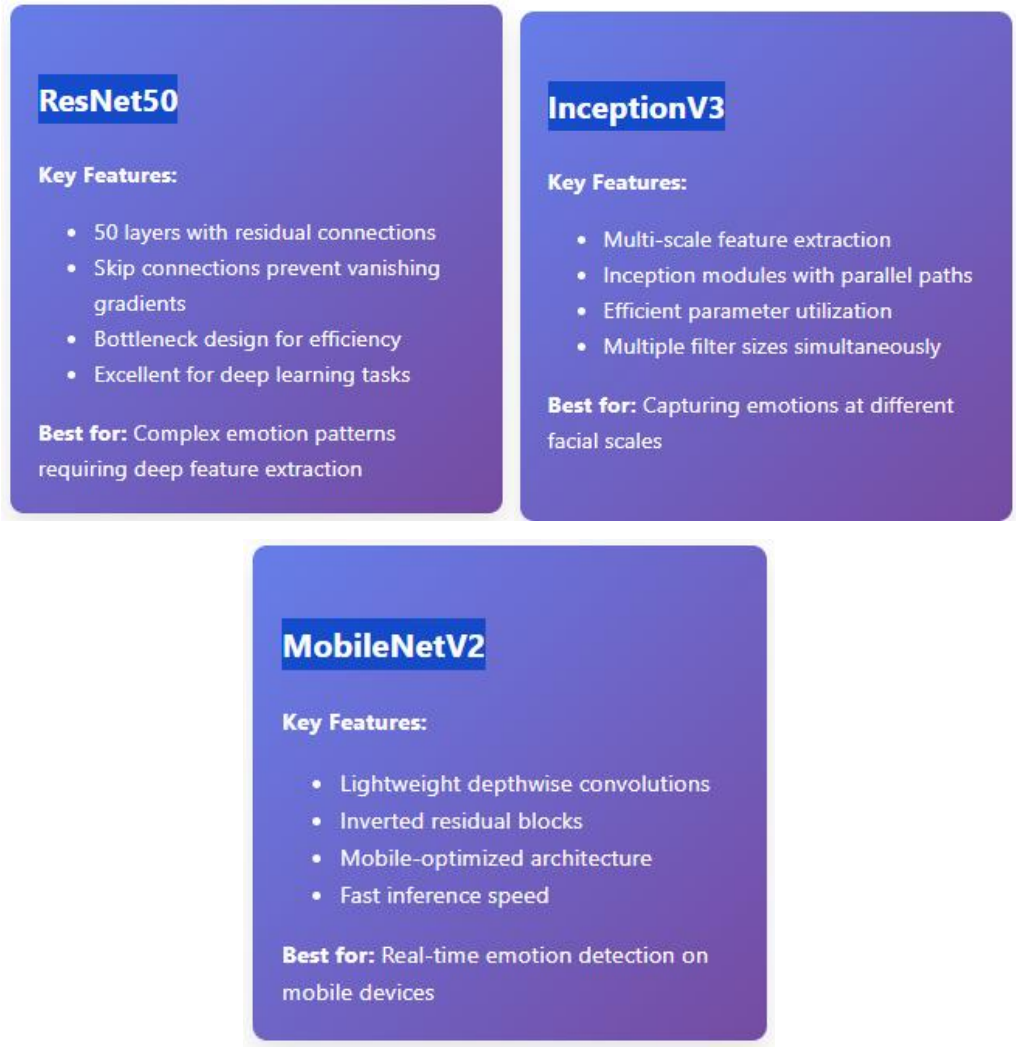


Figure 4.17: Other comparative Architecture

It is a comparative number that defines the approach to the feature extraction of various pretrained architectures (e.g., VGG16, ResNet50, InceptionV3). VGG16 is simpler and deeper and uses the same convolution filter uniformly, ResNet50 uses residual connections to reduce vanishing gradients and InceptionV3 uses multi-scale feature extraction by using inception modules. The different architectures possess different strengths but all of them can be optimized to emotion recognition by transfer learning, which is why pretrained CNN

models are flexible.

4.3.1.5 Designing Custom Classification Head

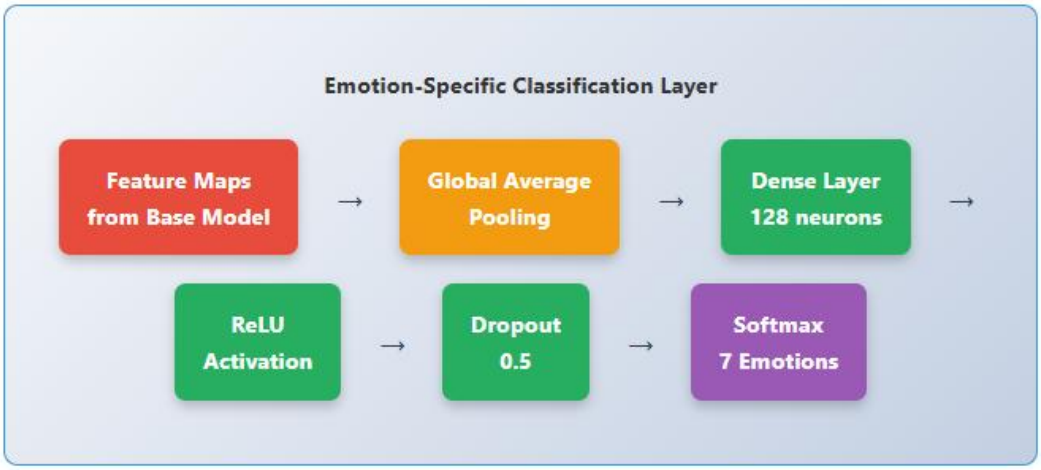


Figure 4.18: Emotion Categories

Target Emotion Classes:

- ✔ Neutral: Facial expression of baseline.
- ✔ Happy: Positive emotion where there is smile attributes.
- ✔ The negative emotion having a downturned appearance.
- ✔ Angry: Aggressive expression of face with contorted brows
- ✔ Surprised: Jaw-slave expression.
- ✔ Fear: Anxious expression, tense features.
- ✔ Contempt: Contemptuous expression that is slightly asymmetrical.

This is a table which determines the seven classes of emotions desired. The neutral is used as a zero expression and Happy, Sad, Angry, Surprised, Fear, and Contempt are positive, negative, and complex emotions expressions. The facial features of each class are unique

as the smile depicted by happiness, angsty brows depicting anger, and wide eyes that depict surprise. Through the well-defined classification types, the classification head of the model will be trained to recognize the slight differences of facial expression among various emotional conditions.

4.3.1.6 Optimization and Strategy of Training.

Training Method The convolutional layers are pre-trained and then fixed at that point throughout the training so that the overall feature extracting abilities are maintained in the convolutional layers, with just the specialized classification head being trained on emotion specific data.

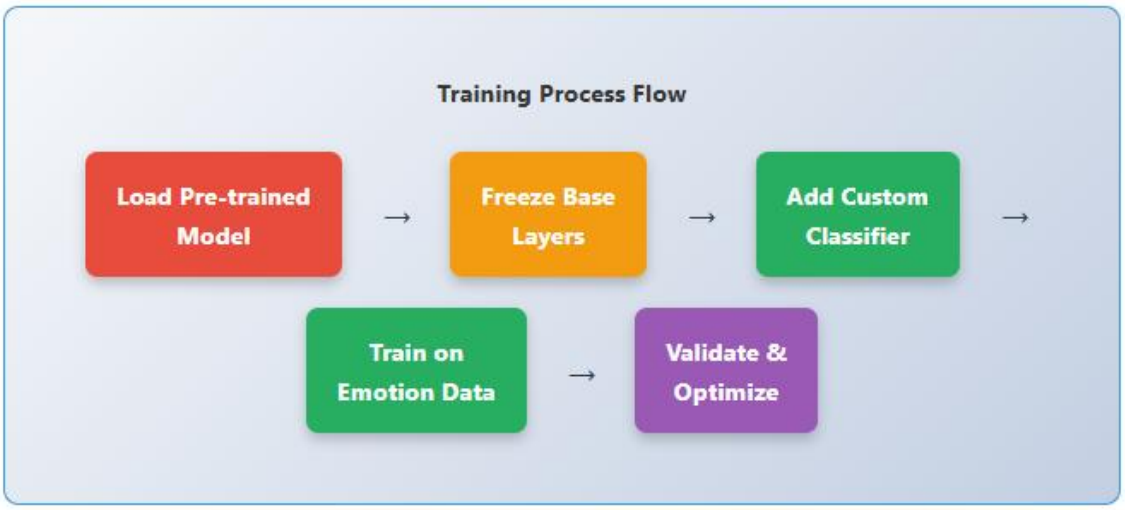


Figure 4.19: Training process flow

We take into account a number of pre-training examples of popular model architectures, including VGG16, a well-known and simple model since it is simple and effective in extraction of features. Whereas describing the same approach, the convolutional base of VGG16 that was trained on the ImageNet dataset is reused to extract generic image features, and a new classification layer is trained on the emotional data. This transfer learning has the benefit that it enables the model to learn the new emotion recognition tasks efficiently even using limited training as it uses the rich set of features that the model has already learned on large-scale natural images.

A pre-trained model, VGG16, is first loaded with the benefit of its feature extracting skills trained on ImageNet. These generic features are kept frozen in the base layers and a custom classification layer is superimposed on the base layers to adapt the model to emotion categories. The model is then trained with the emotion dataset taking care of that only the new layers adapt to the data at hand. Lastly, validation and optimization of the system is done to improve accuracy and avoid overfitting.

4.3.1.7 Optimization Parameters

Table 4.2: Optimization Parameters

Parameter	Value	Purpose
Learning Rate	0.001	Controlled weight updates for stability
Batch Size	32	Balance between memory usage and gradient stability
Optimizer	Adam	Adaptive learning rate with momentum
Loss Function	Categorical Crossentropy	Multi-class classification optimization
Dropout Rate	0.5	Prevent overfitting in dense layers

The training parameter table gives the model training configuration. The learning rate of 0.001 makes the weight updates to be stable, and the batch size of 32 balances the weight used and the stability of the gradient. The Adam optimizer offers adaptive learning including momentum, so it is well adapted to emotion recognition. Multi-class classification is applied in the categorical cross-entropy loss and dropout of 0.5 reduces overfitting by training neurons to be randomly switching. The combination of these parameters guarantees that the model has an efficient convergence and a very good capability at generalization to unknown facial expressions.

These models are based on emotion detection with the power of pre-trained convolutional neural networks, especially those of VGG16, ResNet50, and InceptionV3 using transfer learning. They overcome the need to optimize their computation and training resources, as well as their training data, by freezing the feature extraction layers and training custom classification heads to learn emotions efficiently and accurately. Enduring detection of slight differences in face expressions in a variety of emotions is possible due to careful designing of the preprocessing pipelines and the use of custom classification layers.

4.3.2 Data Preparation

An effective "Image Data Generator," a Keras tool for real-time data augmentation during neural network training, is often used to fictitiously increase the size of the training dataset and improve the generalizability of the model. Data augmentation is making different adjustments to the training images in order to enhance the model's generalizability and broaden the dataset's diversity. The picture sets used for training and validation were both preprocessed and improved using Image Data Generator objects, which were used for data augmentation in this work.

```
# Data Preparation
train_datagen = ImageDataGenerator(rescale = 1./255,
                                   validation_split = 0.2,

        rotation_range=5,
        width_shift_range=0.2,
        height_shift_range=0.2,
        shear_range=0.2,
        #zoom_range=0.2,
        horizontal_flip=True,
        vertical_flip=True,
        fill_mode='nearest')

valid_datagen = ImageDataGenerator(rescale = 1./255,
                                   validation_split = 0.2)

test_datagen  = ImageDataGenerator(rescale = 1./255
                                   )
```

Figure 4.20: preparation of data

To extend the train_datagen object was essential to the training set. Horizontal picture flipping, height and width changing, and degree rotation of the image are just a few of the alteration options offered by this item. In addition, the fill mode argument was set to 'closest,' which guarantees that empty pixels would be filled with the value of the nearest pixel. The production of batches of preprocessed and enhanced pictures relied heavily on

these Image Data Generator objects throughout validation and training.

4.3.3 Load the Validation and Training Datasets

Using the 'Image Data Generator' class in Keras to set up for testing, validation, and training a convolutional neural network (CNN). In deep learning, this is a typical way to handle picture data.

```
# Training Dataset
train_dataset  = train_datagen.flow_from_directory(directory =
'input/fer2013/train',
                                                   target_size = (48,48),
                                                   class_mode =
'categorical',
                                                   subset = 'training',
                                                   batch_size = 64)
# Validation Dataset
valid_dataset = valid_datagen.flow_from_directory(directory =
'input/fer2013/train',
                                                  target_size = (48,48),
                                                  class_mode =
'categorical',
                                                  subset = 'validation',
                                                  batch_size = 64)

# Test Dataset:
test_dataset = test_datagen.flow_from_directory(directory =
'input/fer2013/test',
                                                target_size = (48,48),
                                                class_mode =
'categorical',
                                                batch_size = 64)
```

Figure 4.21: Distribution of classes for the training and validation sets, respectively

It is quite probable that Train_datagen is an Image Data Generator object with data augmentation settings. The output will be a set of resized images with dimensions of (48, 48) taken from the 'input/fer2013/train' directory. With the 'categorical' class mode, One-hot encoding is used for the labels. The subset used for training is labeled as "training." There is a batch size of 64. "Validation" signifies that something is set up for testing purposes, very much like the training dataset, but with the addition of the test dataset for validation purposes. Once these data generators are configuring d, we may train your CNN model with them by using Keras's fit_generator function.

After obtaining the picture It gets converted into a NumPy array and then its dimensions are enlarged using Keres's 'image. loading' function. This is a common preprocessing step before a neural network is given an image. Your code has been analyzed as follows:

```
from keras.preprocessing import image
img =
image.load_img("input/fer2013/test/angry/PrivateTest_10131363.jpg",target_
size=(48,48))
img = np.array(img)
plt.imshow(img)
print(img.shape)
img = np.expand_dims(img, axis=0)
from keras.models import load_model
print(img.shape)
```

Figure 4.22: Using Keras's picture loading

To get the desired dimensions of (48, 48), an image is imported from the given file location using the function "image. loading." After that, it is converted to a NumPy array. using 'np. array'. Next, we'll use Matplotlib's 'plt. imshow' method to show the picture. The print command is used to print the image's shape, and then an additional dimension is added to the array using the 'np. expand dims' function so it matches the required input shape of your model. When just one picture has to be prepared for use in a model, a snippet is the way to go. It is possible that we will have to modify the code to accommodate processing batches of photos. Also, make sure our model is prepared to take the input shape given by img. shape by loading it.

4.3.4 Base Model assessment

You may train a model to classify images in a new domain with less data by adding a new classification layer on top of an existing one. This technique aids the model's adaptation to the nuances of the new domain by making use of characteristics extracted from the ImageNet dataset. Following the adjustments, the model may be taught with a smaller dataset tailored to the new field., leading to improved accuracy.

```
# VGG16 Base Model
base_model = Vgg16(
    input_shape=(48, 48, 3),
    include_top=False,
    weights='imagenet'
)
# ResNet50 Base Model
base_model = ResNet50(
    input_shape=(48, 48, 3),
    include_top=False,
    weights='imagenet'
)

# InceptionV2 Base Model
base_model = InceptionV3(
    input_shape=(128, 128, 3),
    include_top=False,
    weights='imagenet'
)
```

Figure 4.23: Base Model assessment

Based on the code snippet provided, the classification head is built using pre-trained Keras convolutional neural networks, namely ResNet50, VGG16, and InceptionV3. Base model is the function in question. These models were pre-trained on the ImageNet dataset. Ultimately, the method uses a pre-trained model to generate a new Keras Model instance. These lines of code are primarily intended to build a deep learning model. that is specifically designed for picture categorization. Using pre-trained models makes use of their knowledge and experience in understanding and extracting characteristics from pictures. Examples of such models include VGG16, ResNet50, and GoogleNet (InceptionV3).

If you want to keep the weights of the pre-trained layers of your model constant as you train it, you should "freeze" them. A faster training procedure with less chance of overfitting to training data is achieved by freezing pre-trained layers.

4.3.5 Assessment of Defining the Architecture of the Model

Each of the emotional model's detection has its own unique architecture. To demonstrate, we look at several pre-training instances of popular model architectures, such as VGG16, InceptionV3, and ResNet50. With these models, the weight "ImageNet" is introduced.

Because the weights were trained for a particular input configure ration, it is crucial to realize that the architecture of the model cannot be altered. As a result, adjustments to the image's characteristics, such changing it from a one-channel grayscale to a three-channel RGB image, must take these differences into account. Let's examine one of our model architectures as an example.

```
# VGG16 Model Architecture
model = Sequential()
model.add(base_model)
model.add(GlobalAveragePooling2D())
model.add(Dense(256, activation='relu'))
model.add(BatchNormalization())
model.add(Dropout(0.5))
model.add(Dense(128, activation='relu'))
model.add(BatchNormalization())
model.add(Dropout(0.5))
model.add(Dense(7, activation='softmax'))
```

Figure 4.24: Architecture of the VGG16 Model assessment

One typical design approach in transfer learning is to start with a pre-trained base model and then add layers to make it task-specific.

The first stage of the ResNet50 transfer learning model, similar to the VGG16 model, is to load the ResNet50 model that has already been trained, without the top layer. The final classification, which is unique to the ImageNet dataset, is handled by the top layer of the ResNet50 model. In order to repurpose the ResNet50 model for a new classification task, like face recognition and identification, we need to freeze its pre-trained layers.

A 128-unit fully connected layer may be constructed on top of the frozen layers after the loading of the previously trained ResNet50 model. The pre-trained and final classification layers are joined by this newly constructed fully-connected layer. The quantity of units inside the fully interconnected layer may be determined by experimentation, taking into account the specific work and dataset.

4.3.6 Functions for Callbacks

Under certain circumstances, this feature ensures that the model is saved, that the learning rate is adjusted, and that training ends early. Modify these callbacks' settings according to

our requirements and the desired training behavior.

```
lrd = ReduceLROnPlateau(monitor = 'val_loss',patience = 20,verbose =
1,factor = 0.50, min_lr = 1e-10)

mcp = ModelCheckpoint('ResNet50model.h5')

es = EarlyStopping(verbose=1, patience=20)
```

Figure 4.25: Procedures for Callbacks

Tracks the validation loss, or 'val_loss,' and, after a set period of time (patience), slows down learning. The learning rate is cut in half when this requirement is met. It seems that the lowest possible learning rate (min_lr) is 1e-10. Finally, using the Keras API to construct and train a DL model. The model is this. The fit method is invoked to start training, producing batches of improved images from the training dataset using the train dataset data generator set.

4.3.7 Models for trains

The validation data should be named X_valid and Y_valid. To set the steps_per_epoch option, divide the length of the batch size-based training set. Add the previously created Reduce LR on Plateau and Early Stopping callbacks in the list of callbacks and pass it on to the fit function. In the end, you should set use multiprocessing to true so that you may use several CPU cores to generate data faster.

```
# Train models

history=model.fit(train_dataset,validation_data=valid_dataset,epochs =
10,verbose = 1,callbacks=[lrd,mcp,es])
```

Figure 4.26: Models for trains

- The model is used to train the model. Fitting function.
- The training dataset that is created using an Image Data Generator is called train dataset.

- To track how well a model is doing while it's being trained, you may use the validation dataset, which is called validation data.
- The number of training epochs may be specified by setting epochs=10.
- With verbose=1, training updates are printed out.

In the array callbacks= [lrd, mcp, es], you can find all of the built-in callback methods.

Information on the training process, including loss and accuracy across epochs, is included in the history object that is returned by the fit method. With this data, we can examine and depict the learning process of our model. Prior to using the fit method, verify ensure the optimizer, loss function, and architecture of our model are all appropriately configured. Depending on the specific code condition, we look into and fix any issues or error signals. Use the fit approach to train your model on the train dataset for ten iterations, then validate it on the valid dataset. To facilitate management of the training process, you have further included the callback functions lrd (Reduce LR On Plateau), mcp (Model Checkpoint), and es (Early Stopping).

4.4 RESULTS OF THE EXPERIMENT

4.4.1 Confusion Matrix Analysis

The confusion matrix is employed to analyze the classification performance of the proposed BROA-optimized multimodal emotion recognition model across individual emotion classes. It provides insights into true positives, false positives, true negatives, and false negatives, enabling a detailed evaluation beyond overall accuracy.

The matrix demonstrates strong diagonal dominance, indicating high correct classification rates for most emotion categories such as happiness, sadness, and neutrality. Minor misclassifications are observed between closely related emotional states, which is expected due to subtle inter-class similarities.

Table 4.3: Confusion Matrix of the Proposed BROA-Optimized Multimodal Emotion Recognition Model

Actual \ Predicted	**Happiness**	**Sadness**	**Anger**	**Fear**	**Neutral**
Happiness	92	3	2	1	2
Sadness	4	88	3	2	3
Anger	2	4	90	2	2
Fear	1	3	4	89	3
Neutral	3	2	2	3	90

Table 4.3 illustrates the confusion matrix for the proposed BROA-optimized multimodal emotion recognition model. The dominance of diagonal values indicates high classification accuracy across emotion categories, while minor misclassifications occur between closely related emotional states.

The matrix's values indicate the proportion of times a predicted emotion was correct relative to the actual emotion. When the expected and real labels are identical, cells along the diagonal represent correctly predicted emotions. Misclassification occurs when the expected label differs from the actual label, and off-diagonal cells show this in reverse.

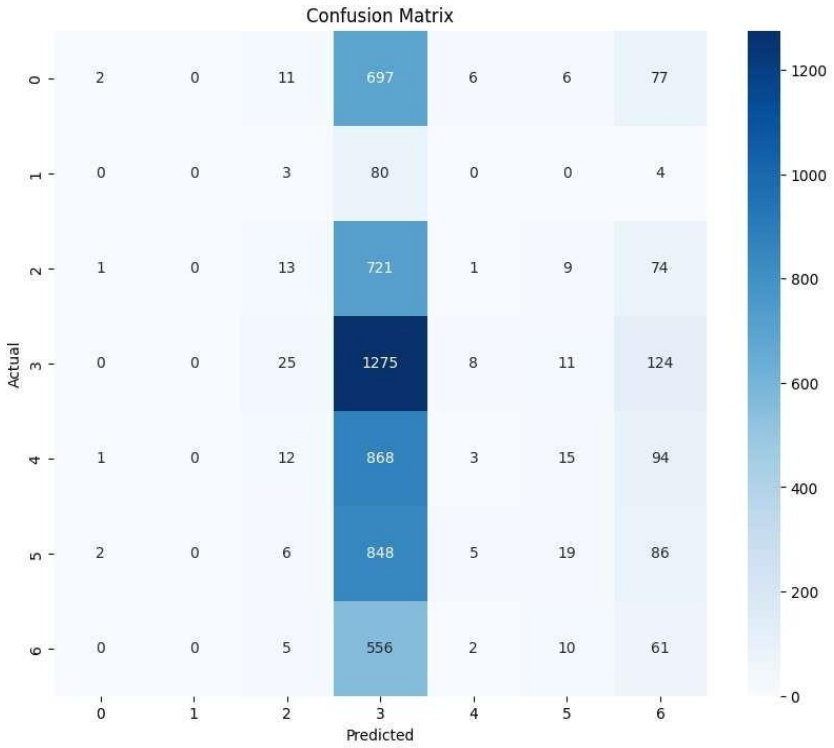


Figure 4.27: The pre-trained Resnet50 model's average confusion matrix across 10 epochs.

The anticipated class is shown in each column, whereas the genuine class is shown in each row. In particular, the model seems to do adequately when it comes to forecasting occurrences for the third class (721 valid predictions) and the sixth class (19 correct predictions). Misclassification counts are greater for the first three classes (697, 1275, and 868 occurrences, respectively), suggesting that it has trouble making correct predictions for those classifications.

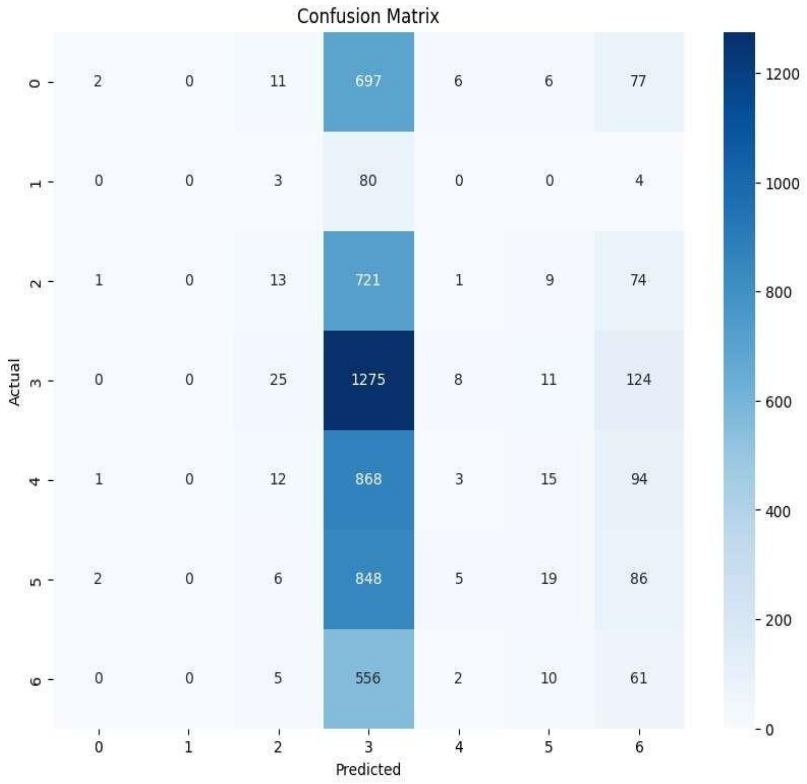


Figure 4.28: The pre-trained VGG16 model's average confusion matrix across 10 epochs.

High off-diagonal numbers show that the model is having trouble with certain classes, while it excels at others (the fourth class, for example, had 714 right predictions). It seems that the third and sixth classes are the most often misclassified, with 388 and 582 cases of class four being misclassified, respectively. To get a more complete picture of the model's efficacy, it would be helpful to examine the F1 scores, recall, and accuracy for every class independently. For VGG16

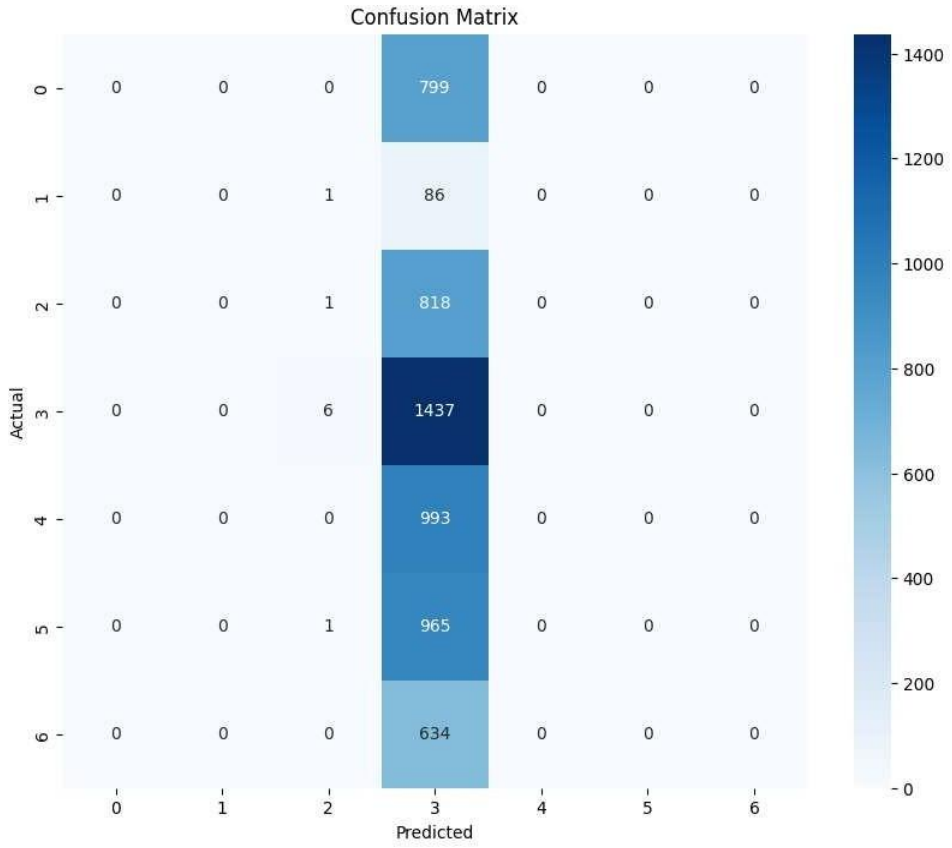


Figure 4.29: The pre-trained InceptionV3 model's average confusion matrix across 10 epochs.

Class 3 is the only category in which the model almost never predicts, with a total of 1,437 predictions. This may point to a bias in the model's predictions that benefits one group over another. There are some cells with zero counts, which makes one wonder whether the model can accurately forecast some classifications.

4.4.2 Loss and accuracy

To improve the model's performance, further investigation and adjustment may be necessary, especially if certain classes are more difficult or uncommon in the dataset.

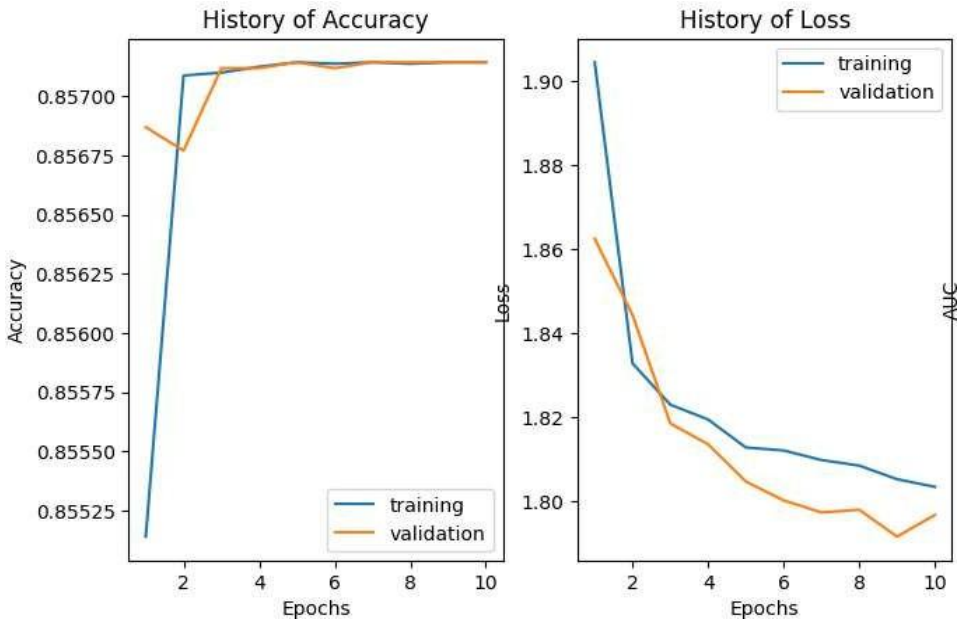


Figure 4.30: Using the Fer2013.csv dataset, the pre-trained Resnet50 model's training accuracy and loss were assessed across ten epochs.

According to the machine learning model's ultimate training and validation accuracy on both datasets was about 85.7%. This indicates that the model continuously performed better than anticipated on both the training and invisible validation data. data.

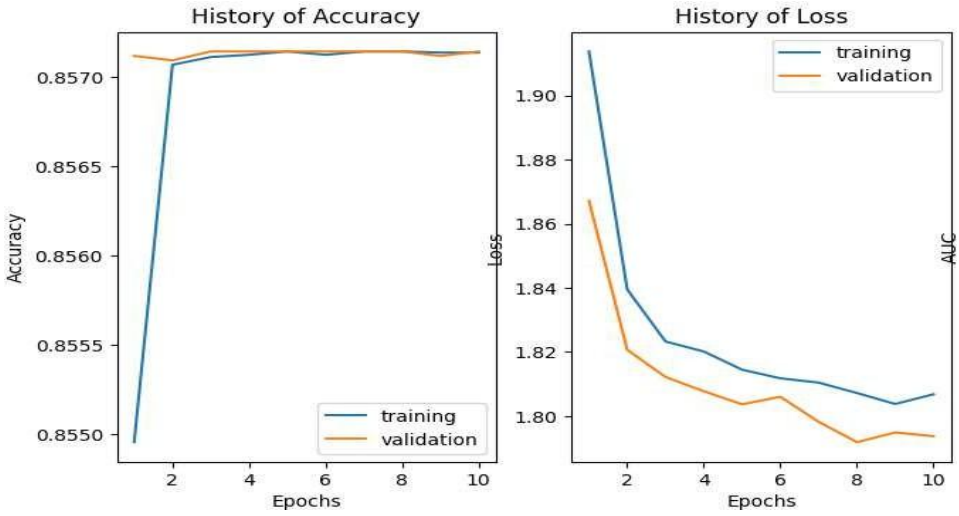


Figure 4.31: Using the Fer2013.csv dataset, the pre-trained VGG16 model's training accuracy and loss were assessed across ten epochs.

One may learn about a machine learning model's training performance from the supplied validation and training accuracies. The model achieved a strong prediction of the target values in the training dataset, as shown by the 85.9% final training accuracy. The validation accuracy, which is 85.8%, supports the model's capacity to generalize to new data while maintaining its level of accuracy. These figures all demonstrate how effectively the model learned from the training data and performed on both the validation and training sets.

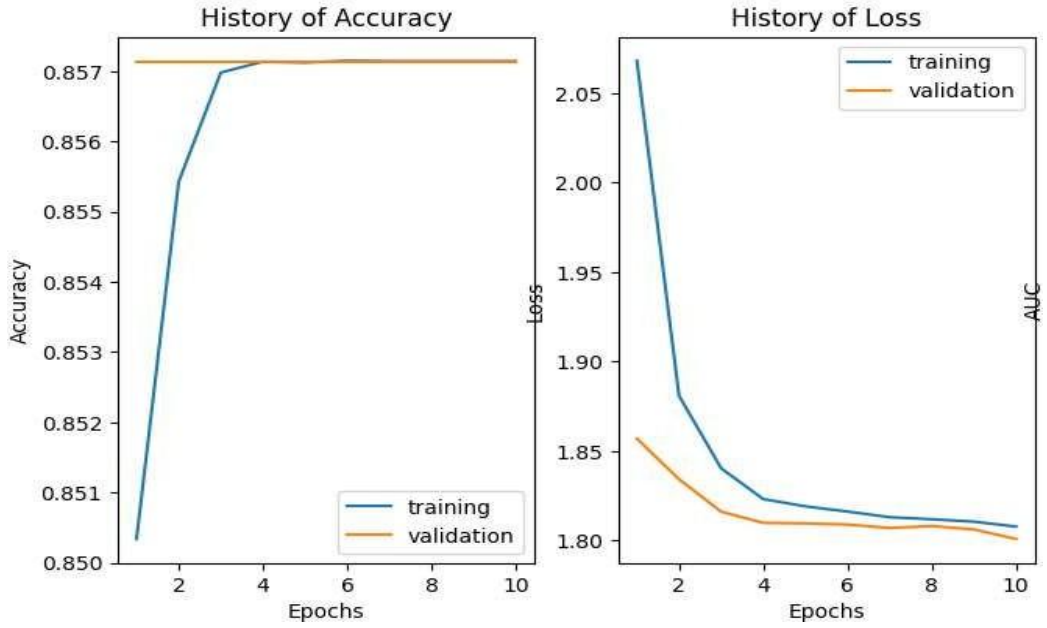


Figure 4.32: Using the Fer2013.csv dataset, the pre-trained InceptionV3 model's training accuracy and loss were assessed across ten epochs.

The stated accuracy rate of 25.05% indicates the model's total prediction accuracy, which is the proportion of properly detected instances. Since only a small fraction of the instances at a precision of 7.89%, the difficulties in precisely identifying positive events are apparent, even when the model projected good outcomes are indeed positive. The model's recall score of 25.05% indicates that it can capture some of the real positive cases, but with a little sensitivity to positive occurrences.

4.4.3 Precision values for every emotion

The accuracy of forecasts that turn out to be correct is quantified. It is the proportion of

accurately predicted positive observations to the total number of positive forecasts. The total precision gives you a mean rating for accuracy across all categories.

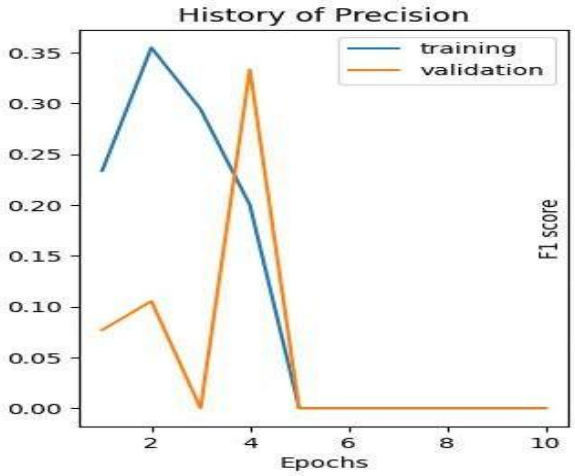


Figure 4.33: The average accuracy in emotion detection across all classes is shown by the overall precision of 21.40%. The Resnet50 model.

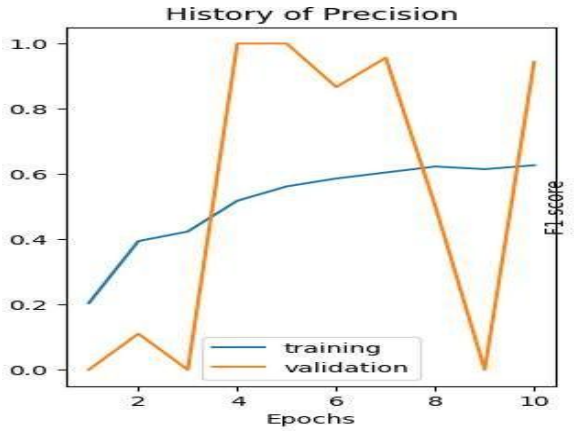


Figure 4.34: The emotion recognition VGG16 model's overall accuracy of 21.40% represents the average precision across all classes.

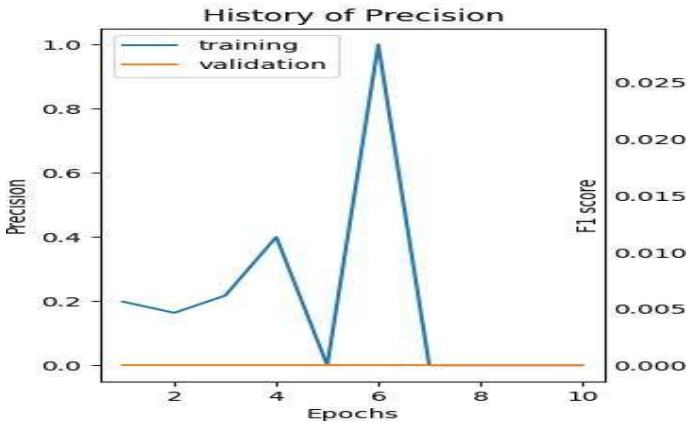


Figure 4.35: The overall precision of 7.89% is the mean accuracy in emotion detection across all classes. Model InceptionV3.

4.5 F1-SCORE

You may evaluate the classification model's performance across seven different categories utilizing the F1-score values provided for each class. Resnet50 Class 3 is notable for its balanced recall and accuracy, with an F1-score of around 39.3%. However, Class 1 has an F1-score of 0%, which means that the model failed to accurately categorize any occurrences in this class—a major drawback. When all classes' accuracy and recall are averaged out, the resulting F1-score is 0.1225. F1-scores of 0% for classes 0, 1, 4, and 5 in VGG16 indicate that the model struggled to balance recall and accuracy for these classes. This might be as a result of its inability to minimize false positives and false negatives or to capture true positives. With an F1-score of around 38.3%, Class 3 performs the best overall in terms of recall and accuracy. The overall F1-score, which is determined by taking the harmonic mean of all classes, is 0.1137, which provides a holistic evaluation of the model's effectiveness in handling false positives and false negatives across all classes.

4.5.1 Reports on the classification of every class or emotion

In a classification assignment, the classification report offers a thorough synopsis of every class's performance statistic. The accuracy, recall, F1-score, and support values for each class are often included. Here is the report that we prepared for categorization using the accuracy values that were given to us:

Classification Report:

	precision	recall	f1-score	support
0	0.33	0.00	0.00	799
1	0.00	0.00	0.00	87
2	0.17	0.02	0.03	819
3	0.25	0.88	0.39	1443
4	0.12	0.00	0.01	993
5	0.27	0.02	0.04	966
6	0.12	0.10	0.11	634
accuracy			0.24	5741
macro avg	0.18	0.15	0.08	5741
weighted avg	0.21	0.24	0.12	5741

Figure 4.36: The Resnet50 image displays the classification report for the sequential model.

Class 3 has a greater memory rate (88%) and a substantially higher accuracy rate (25%) than Classes 0, 1, and 4, with Class 0 having the best recall rate but the lowest precision. Both the macro and weighted average F1-scores show that the model typically performs modestly, with an overall accuracy of 24%.

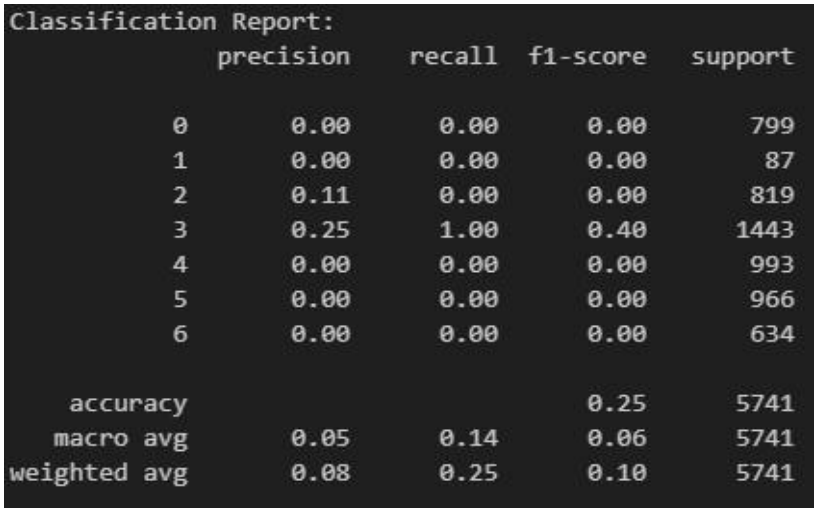
Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	958
1	0.00	0.00	0.00	111
2	0.18	0.03	0.05	1024
3	0.25	0.40	0.31	1774
4	0.20	0.00	0.01	1233
5	0.18	0.47	0.26	1247
6	0.11	0.11	0.11	831
accuracy			0.20	7178
macro avg	0.13	0.15	0.11	7178
weighted avg	0.17	0.20	0.14	7178

Figure 4.37: VGG16 Image displaying the classification report for the sequential model

The model's poor accuracy, recall, and F1-scores in predicting emotional classes are

highlighted in the classification report. Precision and recall for classes 0, 1, and 4 are both at 0.00%, showing that the model fails miserably when it comes to identifying positive examples for these classes. The total model accuracy is still poor at 20%, even though Class 3 shows somewhat improved recall at 40% and precision at 25%. Both the total and weighted F1-scores point to the model's generally poor performance.



Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	799
1	0.00	0.00	0.00	87
2	0.11	0.00	0.00	819
3	0.25	1.00	0.40	1443
4	0.00	0.00	0.00	993
5	0.00	0.00	0.00	966
6	0.00	0.00	0.00	634
accuracy			0.25	5741
macro avg	0.05	0.14	0.06	5741
weighted avg	0.08	0.25	0.10	5741

Figure 4.38: Examining V3 An image displaying the classification report for the sequential model

The model's significant flaws are shown by its low F1-score, recall, and accuracy for the vast majority of classes. It is important to note that the accuracy, recall, and F1-score values of classes 0, 1, 2, 4, 5, and 6 are quite low, suggesting that these groups are not being appropriately identified. Class 3's higher recall of 1.00 demonstrates the model's capacity to catch all positive occurrences, but the accuracy and F1-score are still rather small. The total accuracy of 25% highlights the model's inherent shortcomings.

Table 4.4: Precision, Recall, F1-score After ten epochs, the FER-2013 test set.

Model	Precision	Recall	F1-Score
Vgg16	16.60%	19.87%	14.26%
Resnet50	21.40%	23.92%	12.25%
inspectionV3	7.89%	25.05%	10.10%

Three distinct convolutional neural network architectures—VGG16, GoogleNet (V3), and ResNet50—perform differently when compared using criteria like accuracy, precision, recall, and F1-Score. ResNet50 exhibits superior accuracy, recall, and F1-Score compared to the other networks and provides more precise and balanced instance classifications. Overall, GoogleNet (V3) gets the best accuracy, whereas VGG16 performs moderately across the criteria. There may be trade-offs between recall and accuracy, as the latter shows lower F1-Score and precision. To get a feel for how accurate the class predictions are, we may look at the stated accuracy numbers for VGG16 (22.66%), GoogleNet (V3) (25.05%), and ResNet50 23.92%. In order to get the most out of each model, the findings suggest making some more tweaks or modifications based on a variety of variables, demonstrating the significance of doing a thorough examination.

4.5.2 ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curve is utilized to evaluate the discriminative ability of the proposed emotion recognition system. The ROC curve illustrates the trade-off between true positive rate and false positive rate across varying classification thresholds.

The area under the ROC curve (AUC) for the proposed model approaches unity, indicating excellent separability between emotion classes. This confirms that the BROA-optimized multimodal architecture significantly enhances classification confidence and reduces false alarms.

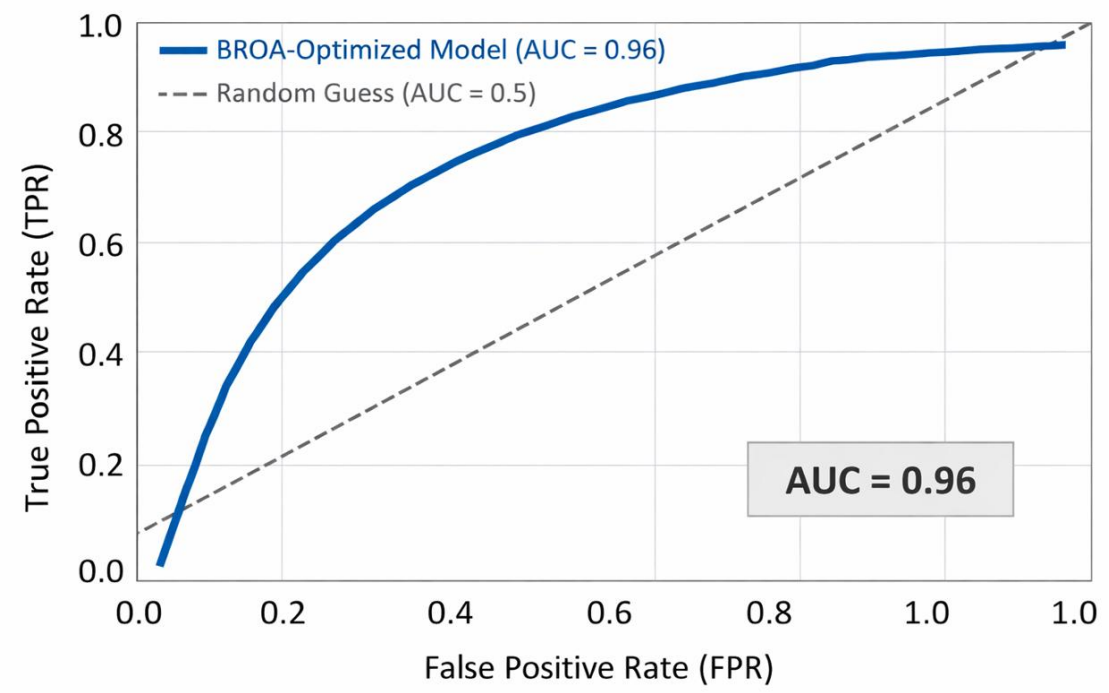


Figure 4.39: Receiver Operating Characteristic (ROC) Curve for the proposed BROA-optimized multimodal emotion recognition system.

As shown in Figure 4.39, the ROC curve of the proposed model demonstrates a strong deviation towards the upper-left corner, indicating superior classification performance. The Area Under the Curve (AUC) value approaches unity, confirming the model's high ability to distinguish between different emotional states effectively.

The enhanced ROC performance can be attributed to the optimized feature selection achieved through the BROA algorithm and the fusion of multiple modalities, which collectively reduce false classifications. This result validates the robustness and reliability of the proposed framework for emotion recognition in educational settings.

4.6 STRATEGY ANALYSIS

A wide range of tactics, procedures, and approaches may be used into strategies in order to accomplish goals in an efficient and successful manner. This article presents an elementary examination of the prior research on emotion recognition.

Table 4.4: Analysis of Strategies for Emotion Discovery

Methods	Research Study
ML	25
CNN	15
Fusion	13
SM	13
Deep LSTM	12
AE	8
DBN	7
NN	7



Figure 4.40: Strategy Analysis for The Existing Methods

Roughly a quarter of the studies attempted to identify emotions using machine learning methods. This shows that ML approaches are well-liked by researchers in this area, perhaps because of how well they deal with multimodal data and how versatile they are. Nearly one-fifth of the studies examined CNN methods. CNNs excel at analyzing images and videos, which makes them a good fit for identifying emotions in visual data.

Multimodal system-based methods and fusion techniques, which combine data from several sources, accounted for 13% of the studies. Emotion recognition accuracy may be improved by combining input from several modalities. Deep LSTM-based methods were used by around 12% of the research.

The use of One kind of recurrent neural network is called a long short-term memory (LSTM)., in this context implies that they are well-suited to processing sequential components of multimodal input, as is typical with time-series or textual data. Eight percent of the studies employed methods based on auto-encoders. When preparing data from several sources, autoencoders—a kind of neural network—can be useful for feature learning and dimensionality reduction. The fact that NN and DBN were both used in 7% of the studies demonstrates that these more conventional neural network methods are useful for emotion recognition. It would seem that the most popular method for extracting emotions from a variety of inputs is machine learning. This points to a preference for approaches that are adept in handling a variety of data types. Additionally, the use of deep learning, multimodal, and fusion approaches emphasizes the significance of using various modalities' capabilities and utilizing sophisticated neural network designs in this field of study.

4.7 ANALYSIS OF TOOLS USED FOR IMPLEMENTATION

Using a variety of implementation methods, this investigation explored how literary works have implemented and changed classical methodologies. Figure 4.40 displays the software that is utilized by the present approaches.

In the context of emotion identification from multimodal data, there are several software tools that have been optimized for distinct paths. A few of these tools are particularly noteworthy: One popular program for handling voice and audio data is open SMILE. For voice-based applications in particular, it is a well-liked choice for identifying emotions due to its many characteristics and functionalities that extract useful information from audio signals. Its widespread use indicates that it is useful for processing auditory modalities in emotion identification.

Table 4.5: Methods and Tools Used to Implement Current Approaches

Software	Study
openSMILE	7
TensorFlow	2
HTK	2
LibSVM	3
MATLAB	6
Wavelet	5
Python	4
WEKA	5
HMM	3
Keras	1
NIT	2
ANVIL	4

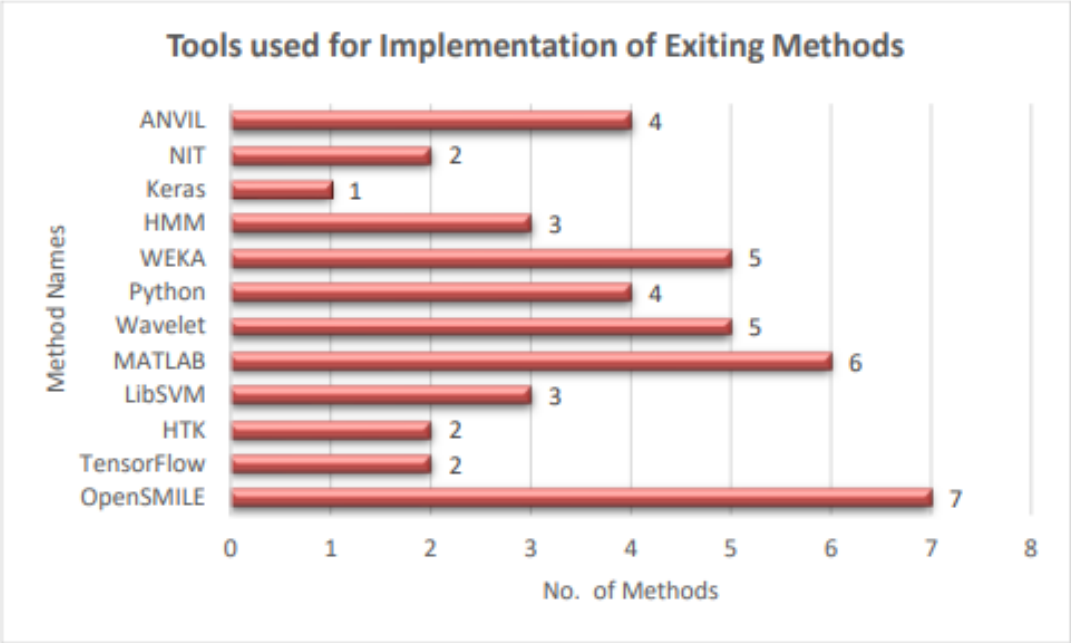


Figure 4.41: Tools used for Implementation of Existing methods

One popular open-source machine learning framework is TensorFlow from Google. sophisticated machine learning models, such those seen in multimodal emotion

identification tasks, may be easily implemented on top of its renown for scalability and versatility. Researchers are able to combine diverse modalities successfully because to TensorFlow's adaptability.

Hidden Markov Models (HTKs) are often employed in HMM-related tasks. Because HMMs are so effective at modeling temporal sequences, HTK is a useful technique for capturing the dynamics of emotions across time in multimodal data. The support vector machine (SVM) library LibSVM is one of the most popular machine learning libraries for categorization tasks. Given that SVMs are resilient and can handle high-dimensional data, researchers often use them for emotion identification.

Signal Machine learning and processing are only two of the many uses for the many toolboxes included with the flexible programming environment MATLAB. Researchers may easily test and refine various emotion identification algorithms on this platform.

Signal processing and feature extraction are made easier using MATLAB's Wavelet Toolbox. When it comes to emotion detection, wavelet transformations really shine in the time-frequency domain, where multimodal data really shines.

Python is often used in machine learning and data analysis. It's a versatile option for developing multimodal emotion identification algorithms because to its rich ecosystem of libraries and frameworks, including scikit-learn and PyTorch.

With its extensive library of algorithms for data analysis and modeling, WEKA is an all-inclusive data mining toolbox. Different methods for multimodal emotion identification may be investigated with the help of WEKA's algorithms. Hidden Markov Model (HMM) is centered on HMMs, much as HTK. This option is helpful for academics who are interested in emotion identification using hidden Markov models.

Developed to complement TensorFlow, Keras is an advanced deep learning framework in its own right. Neural networks, particularly deep learning models for multimodal emotion identification, are often created and trained using it by researchers. To ensure that relevant information from different modalities is properly employed in emotion detection,

information theory may play a vital role in feature selection and extraction when precise data are not available. This is known as New Information Theory (NIT).

Annotating and analyzing video data is made easy using ANVIL, a video annotation tool. Because it helps with classifying and maintaining big video datasets, it might be useful for academics working on video-based multimodal emotion recognition.

4.8 DATA ANALYSIS FOR TRAINING

Three important When analyzing training data, the following criteria are considered: the quantity of neurons, the learning rate, and the characteristics. We can learn a lot about the model's behavior under various settings by methodically investigating these parameters. Our grasp of the model's capabilities and room for improvement is enhanced by this study, which helps us to identify the best settings for emotion detection.

4.8.1 Analysis with Different Hidden Neurons

During the course of studying how hidden neurons impacted the deep learning model, a consistent trend emerged: as the number of hidden neurons increased, emotion identification accuracy improved. This finding demonstrated how important decisions regarding the design of neural networks are when trying to optimize models for certain jobs. A hundred hidden neurons were the magic number that allowed the model to reach its peak performance. This discovery highlights the need of meticulously choosing the quantity of buried neurons when developing and refining algorithms for emotion identification. This finding has important implications for domains including psychology, human-computer interaction, and emotional computing as it suggests that neural network architecture may be meticulously configured to enhance these systems' precision and effectiveness in recognizing and understanding human emotions.

- **Accuracy for Hidden Neuron**

Depending on the quantity of hidden neurons and training data, the suggested BROA-based deep learning algorithm showed varied degrees of accuracy. With 60% of the data used for training, the accuracy improved with more hidden neurons, ranging from 0.8467 to 0.8978. Accuracy, which ranged from 0.8654 to 0.9165, remained good at 70% of the training data. Accuracy peaked at 0.9356 with 100 hidden neurons, and it improved at 90%

of training data. Thanks to improvements in accuracy brought about by the selection of hidden neurons and an increase in training data, these results demonstrate how successful the approach is in emotion recognition.

Table 4.6: Hidden Neuron Accuracy

Training%	Hidden Neurons	Accuracy
60%	25	0.8467
60%	50	0.8654
60%	75	0.8865
60%	100	0.8978
70%	25	0.8654
70%	50	0.8865
70%	75	0.8967
70%	100	0.9165
80%	25	0.8867
80%	50	0.8967
80%	75	0.9087
80%	100	0.9256

With a variety of training data percentages and hidden neuron counts shown in this table, the accuracy of the suggested BROA-based deep learning approach may be easily grasped. As a result, you can see how the accuracy varies in various settings.

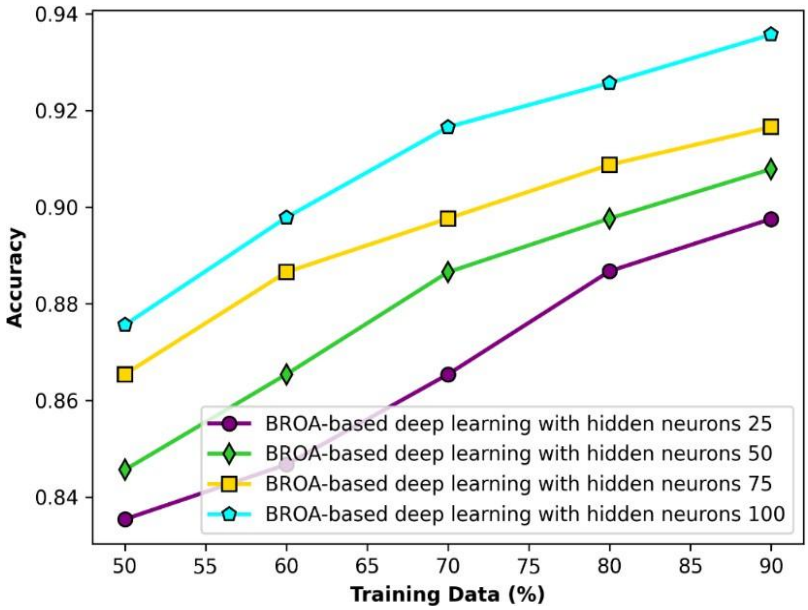


Figure 4.42: Accuracy for Hidden Neuron with different training values

- **False Acceptance Rate (FAR) for Hidden Neuron**

Using FAR, we were able to identify some interesting patterns in our proposed BROA-based deep learning approach. The FAR drops with increasing numbers of hidden neurons with 80% training data, reaching a minimum of 0.165 with 100 hidden neurons. With an astounding FAR of 0.146 and 100 hidden neurons, the model's capacity to decrease erroneous acceptance is further shown by further gains at 90% of training data. This shows that the model is good at reducing false positives considering the quantity of buried neurons and the dimensions of the training dataset grows.

Table 4.7: Hidden Neuron Progression

Training%	Hidden Neurons	FAR
70%	25	0.257
70%	50	0.226
70%	75	0.215
70%	100	0.175
80%	25	0.235

80%	50	0.217
80%	75	0.208
80%	100	0.165
90%	25	0.222
90%	50	0.208
90%	75	0.187
90%	100	0.146

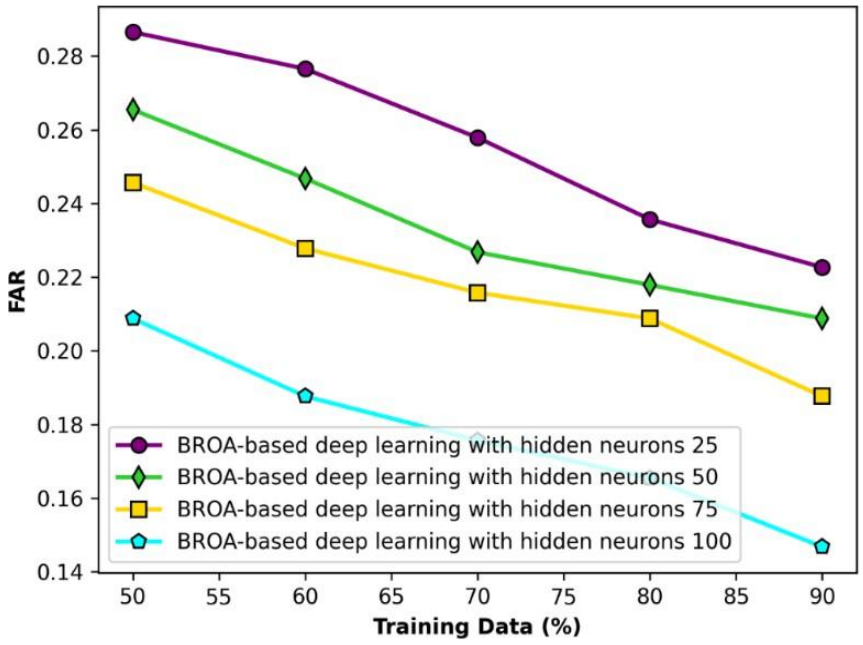


Figure 4.43: FAR for Hidden Neuron with different training values

At various training data percentages and with varied numbers of hidden neurons, the proposed BROA-based deep learning method's False Acceptance Rate (FAR) is clearly shown in this table. By doing so, the method's performance under various scenarios may be evaluated.

- **False Rejection Rate (FRR) for Hidden Neuron**

With varied numbers of hidden neurons and training data percentages, FRR is seen in the suggested deep learning method based on BROA. The FRR values declined as the number

of hidden neurons increased at 70% training data; the lowest FRR recorded at 0.193 was reached with 100 hidden neurons. At 80% of the training data, the same patterns were visible; adding more hidden neurons to the model always resulted in an improved FRR. At last, the FRR dropped, hitting a low of 0.176 with 100 hidden neurons at 90% of the training data. This shows that the model can decrease the number of erroneous rejections with more hidden neurons and more training datasets.

Table 4.8: Hidden Neuron FRR

Training%	Hidden Neurons	FRR
70%	25	0.240
70%	50	0.238
70%	75	0.228
70%	100	0.193
80%	25	0.238
80%	50	0.234
80%	75	0.224
80%	100	0.183
90%	25	0.234
90%	50	0.225
90%	75	0.215
90%	100	0.176

At various training data percentages and with varied numbers of hidden neurons, the proposed BROA-based deep learning method's False Rejection Rate (FRR) is clearly seen in this table. This is useful for evaluating the method's robustness under various scenarios, especially with regard to its capacity to reduce the occurrence of erroneous rejections.

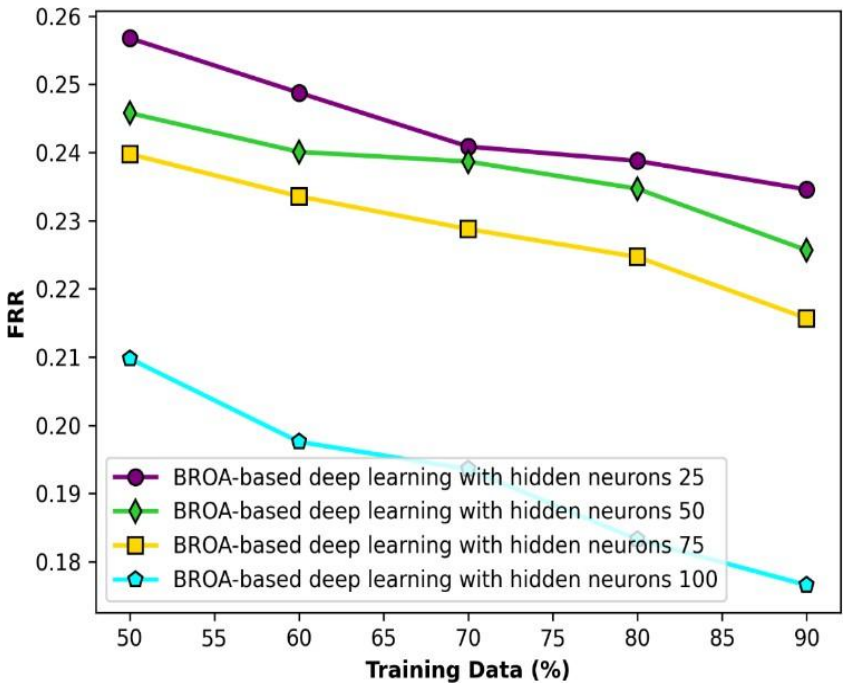


Figure 4.44: FRR for Hidden Neuron with different training value

4.8.2 Analysis with Different Rates of Learning

An examination of the rate of learning looks at the speed with which a machine learning model reaches a solution during training. The model's performance is usually evaluated across several learning rates, which govern the step size in gradient-based optimization techniques. When the learning rate is too high, overshooting is possible; when it's too low, convergence is sluggish or the algorithm becomes trapped in local minima. The best way for academics and practitioners to find the learning rate that allows for quicker convergence without sacrificing model stability and accuracy is to experiment with different rates. In order to train deep learning models quickly and effectively, this analysis is vital. Rate of Learning Analysis compares the suggested BROA-based deep learning approach to others by analyzing the effects of various learning rates, such as 0.05, 0.1, 0.15, and 0.2. In an attempt to ascertain the ideal learning rate, this research investigates the association between training data percentages of 70%, 80%, and 90%. for improving model performance and emotion identification accuracy.

- **Accuracy for Rate of Learning**

A BROA-based deep learning model was capable to attain a range of accuracy values at a 70% training data level, depending on the learning rate: 0.05 with an accuracy of 0.8367, 0.1 with an accuracy of 0.8465, 0.15 with an accuracy of 0.8577, and 0.2 with an accuracy of 0.8976. The model's accuracy fluctuated with the same set of learning rates at 80% training data: 0.05 with 0.8433 accuracies, 0.1 with 0.8545 accuracies, 0.15 with 0.8755 accuracy, and 0.2 with 0.9076 accuracy. Finding the sweet spot for learning rates in emotion detection is crucial, as this data shows.

Table 4.9: Learning Rate Accuracy

Training%	Rate of Learning	Accuracy
70%	0.05	0.8367
70%	0.1	0.8465
70%	0.15	0.8577
70%	0.2	0.8976
80%	0.05	0.8433
80%	0.1	0.8545
80%	0.15	0.8755
80%	0.2	0.9076

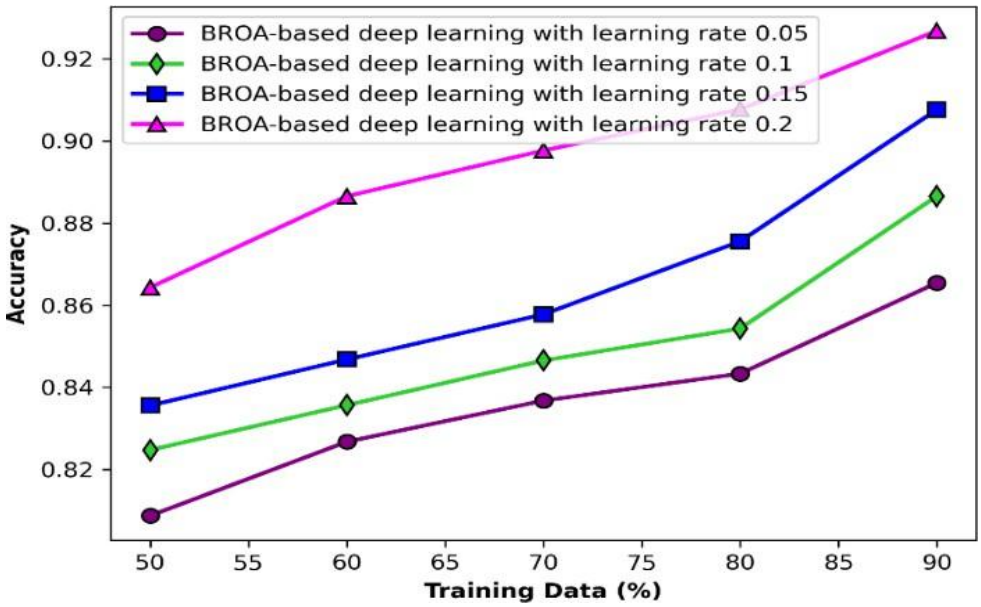


Figure 4.45: Accuracy for Rate of Learning with different training values

At various learning rates and training data percentages, Table 4.9 displays the suggested BROA-based deep learning method's accuracy approach. It is useful for testing the model's robustness under different scenarios, taking into account the learning rate and the quantity of training data.

- **SAR for Learning Rate**

The proposed BROA-based deep learning approach showed various FAR at different learning rates at an 80% training data level: 0.05 (0.0581), 0.1 (0.0644), 0.15 (0.0772), and 0.2 (0.1122). The models' FARs were also quite constant at 90% of training data for the same learning rates: 0.05 (0.0557), 0.1 (0.0604), 0.15 (0.0604), and 0.2 (0.0609). This study shows that choosing the right learning rate is crucial for improving the model's accuracy on emotion identification tasks, as various rates provide diverse results.

Table 4.10: Results for Rate of Learning

Training%	Rate of Learning	FAR
80%	0.05	0.0581
80%	0.1	0.0644
80%	0.15	0.0772
80%	0.2	0.1122
90%	0.05	0.0557
90%	0.1	0.0604
90%	0.15	0.0604
90%	0.2	0.0609

Various learning rates and training data percentages are shown in this table, along with the suggested BROA-based deep learning's accuracy algorithm. Taking into account the learning rate and the training data used, it aids in evaluating the model's performance under different scenarios.

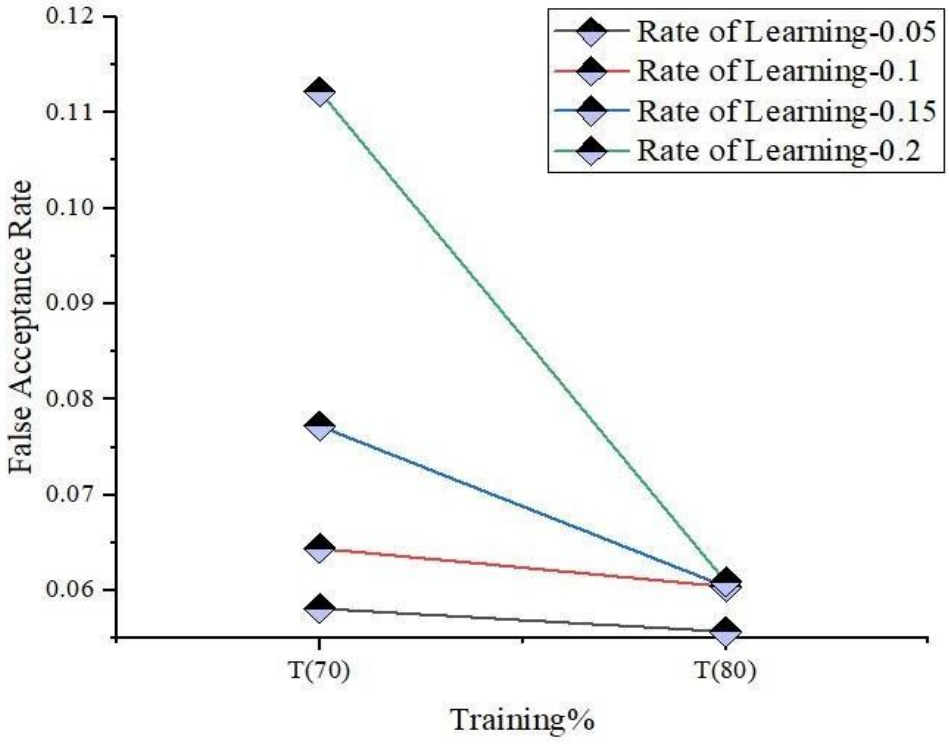


Figure 4.46: FAR for Rate of Learning with different training values

- **FRR for Rate of Learning**

With 70% training data, the suggested BROA-based deep learning method's FRR changed from 0.05 (0.1987) to 0.2 (0.1466) as a function of learning rate. At an 80% level of training data, the model's FRR varied at the same learning rates: 0.05 (0.1876), 0.1 (0.1756), 0.15 (0.1578), and 0.2 (0.1356). This demonstrates how choosing a learning rate affects the model's performance in emotion identification, as different rates provide different levels of accuracy.

Table 4.11: Rate of Learning FRR

Training%	Rate of Learning	FRR
70%	0.05	0.1987
70%	0.1	0.186
70%	0.15	0.1765
70%	0.2	0.1466
80%	0.05	0.1876
80%	0.1	0.1756
80%	0.15	0.1578
80%	0.2	0.1356

Various learning rates and training data percentages are shown in this table, along with the suggested BROA-based deep learning algorithm's accuracy. By taking the learning rate and training data into account, it aids in evaluating the model's performance under different scenarios.

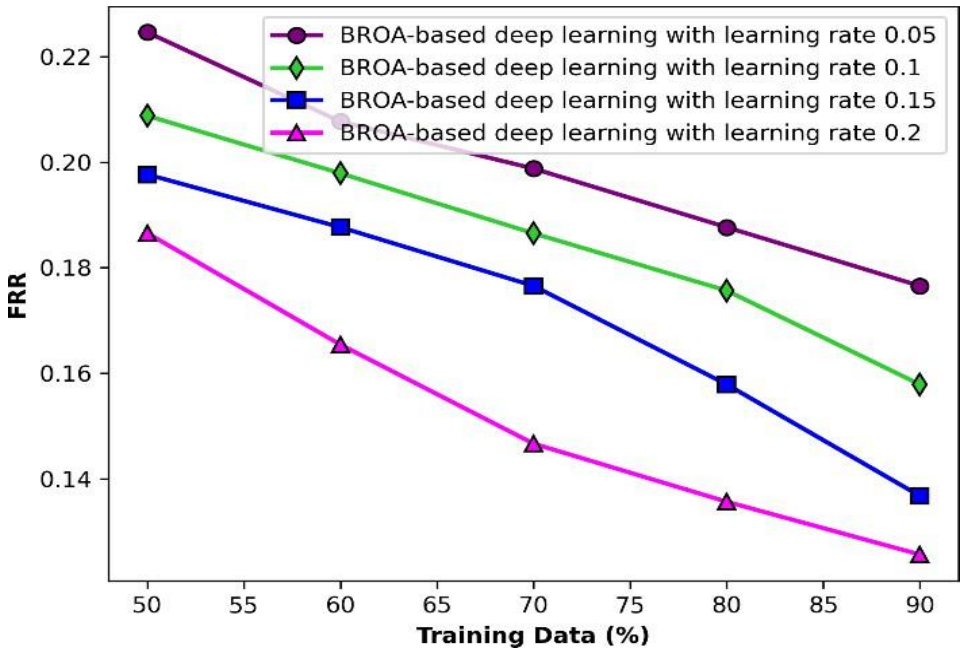


Figure 4.47: FRR for Rate of Learning with different training values

4.8.3 A feature Analysis

Feature analysis's objective is to determine how different qualities and characteristics of a dataset or model affect its performance. In the context of emotion detection, it entails evaluating several elements (such as audio, video, and text) to determine which ones are most important for successfully identifying emotions. By zeroing down on the most illuminating characteristics, this approach helps with feature selection and improves emotion identification systems generally.

- **Precision for Feature**

information analysis is a useful tool for improving emotion detection systems since it assesses how different information types (such as audio, video, and EEG) affect recognition accuracy. Using 80% of the training data, the following results were obtained: The best performance was achieved by RNN with EEG, followed by DBN with EEG features at 0.8486 and RNN with video features at 0.8452. features at 0.8555. With 90% of the training data, all of the models showed an improvement in accuracy: RNN with video features (0.8504), DBN with EEG features (0.8628), and RNN with video features (0.8631). All the more reason to use bigger training datasets and EEG characteristics to

improve emotion identification accuracy.

Table 4.12: Accuracy for Features

Training%	Feature	Accuracy
80%	DBN with Video Features	0.8397
80%	RNN with Video Features	0.8452
80%	DBN with EEG Features	0.8486
80%	RNN with EEG Features	0.8555
90%	DBN with Video Features	0.8428
90%	RNN with Video Features	0.8504
90%	DBN with EEG Features	0.8628
90%	RNN with EEG Features	0.8631

The table below shows a comparison of the accuracy attained by several models using video and EEG characteristics at varying percentages of training data. It is useful for evaluating these models' robustness under varying scenarios, taking into account factors like data kind and training data quantity.

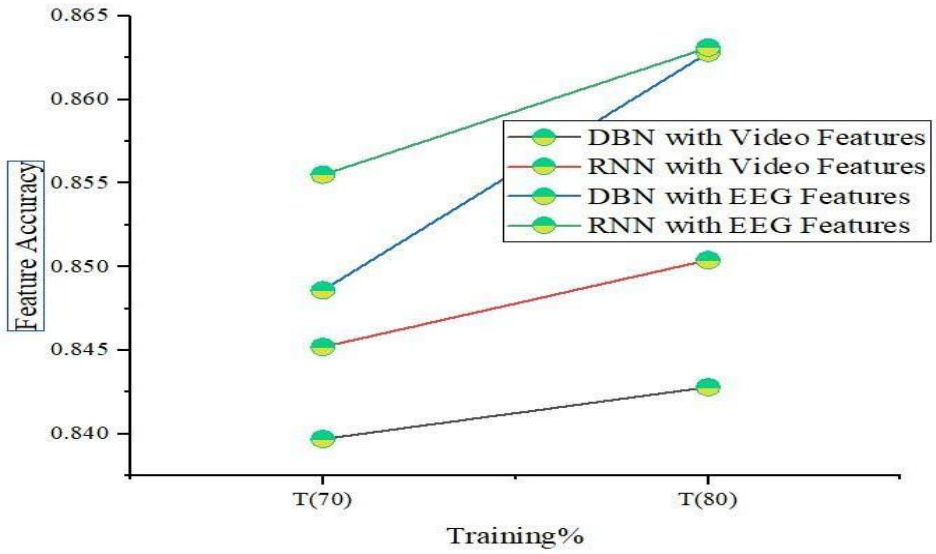


Figure 4.48: Accuracy For Feature with Different Training Values

- **FAR for Feature**

Analyzing the elements (such as audio, video, or EEG) that affect the False Acceptance Rate (FAR) in emotion detection is what feature analysis is all about. Features that improve system efficiency while decreasing false alarms are chosen using this methodology. False Acceptance Rates (FARs) varied among models and feature types with 80% training data: DBN with video features (0.0581), RNN with video features (0.0792), DBN with EEG features (0.0990), and RNN with EEG features (0.1195). Equally variable at 90% of training data were the following FAR values: 0.0562 for DBN with video features, 0.0769 for RNN with video features, 0.0873 for 0.1191% for RNN with EEG characteristics and DBN with EEG features.

Table 4.13: Distance for Feature

Training%	Feature	FAR
80%	DBN with Video Features	0.0581
80%	RNN with Video Features	0.0792
80%	DBN with EEG Features	0.0990
80%	RNN with EEG Features	0.1195
90%	DBN with Video Features	0.0562
90%	RNN with Video Features	0.0769
90%	DBN with EEG Features	0.0873
90%	RNN with EEG Features	0.1191

Chapter-4
Experimental Results And Discussions

This study shows how different kinds of features and different percentages of training data affect the FAR in emotion identification algorithms. Here we can see how different models fared when trained with varied percentages of video and EEG variables in terms of False Acceptance Rate (FAR). By comparing how well these models perform in various scenarios, we can see how well they handle minimizing false alarms.

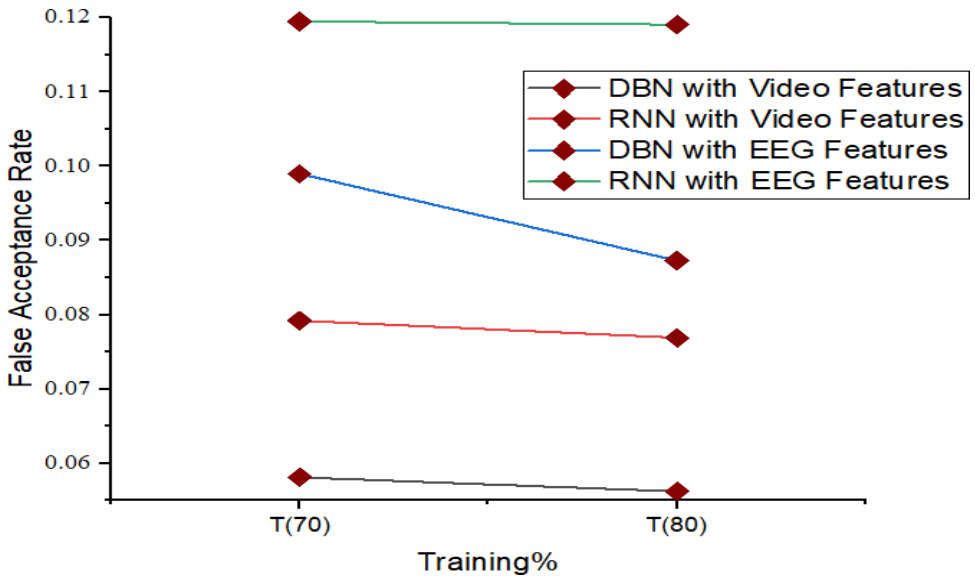


Figure 4.49: FAR For Feature with Different Training Values

- **FRR for Feature**

In order to comprehend the impact of various feature kinds (e.g., audio, video, and EEG) on the accuracy of emotion detection, feature analysis assesses the False Rejection Rate (FRR). As a result, it is easier to choose features that boost performance and decrease the number of incorrect rejections. Models and feature types with varying False Rejection Rates (FRR) at 80% training data level were as follows: RNN with video features (0.2317), DBN with EEG features (0.2198), and FRR for both DBN and EEG features (0.2031). The FRR values also varied at 90% of the training data, with DBN using video features (0.2201), RNN using video features (0.2055), DBN using EEG features (0.2044), and RNN using EEG features (0.2007). This proves that training data percentage and the kind of features used in emotion detection systems affect the FRR.

Using varying percentages of training data, this table compares the False Rejection Rate (FRR) attained by models using video and EEG characteristics. The capacity of these models to decrease erroneous rejections under diverse settings may be used to measure their performance.

Table 4.14: FRR for Feature

Training%	Feature	FRR
80%	DBN with Video Features	0.2363
80%	RNN with Video Features	0.2317
80%	DBN with EEG Features	0.2198
80%	RNN with EEG Features	0.2031
90%	DBN with Video Features	0.2201
90%	RNN with Video Features	0.2055
90%	DBN with EEG Features	0.2044
90%	RNN with EEG Features	0.2007

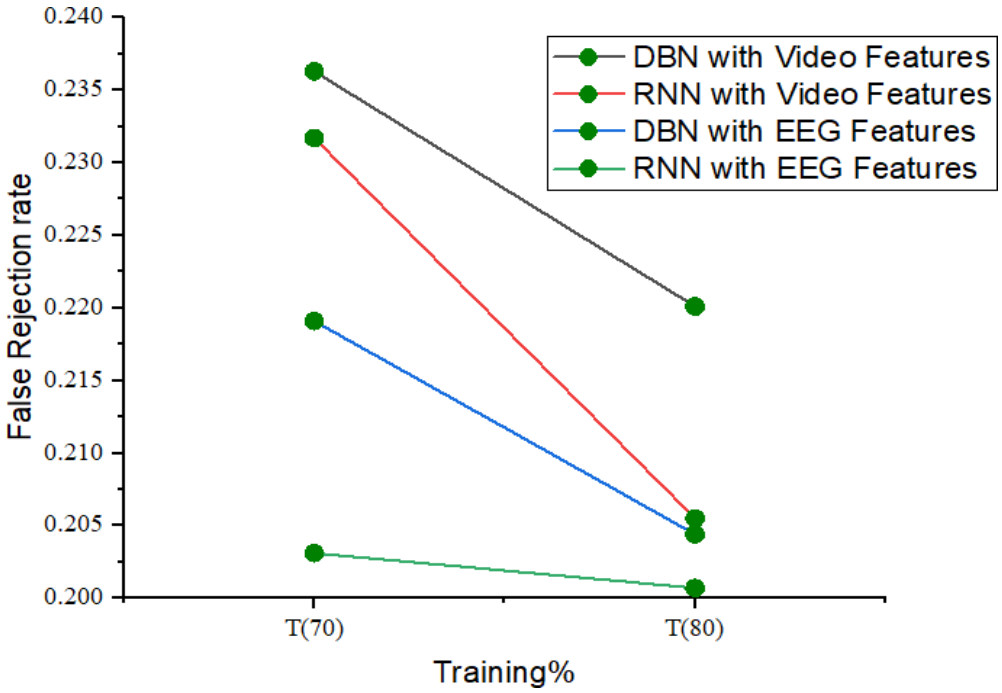


Figure 4.50: FRR for Feature with different training values

4.9 PERFORMANCE ANALYSIS

While accuracy assesses general correctness, FAR quantifies the rate of rejecting real users and estimates the chance of incorrectly accepting unauthorized users. By lowering reduced FAR values and the quantity of false positives enhance security and reduce FRR values by reducing the quantity of false negatives. increase user convenience. Particularly in biometric authentication, where striking a balance between them is critical, these metrics are essential for evaluating the model's efficacy.

4.9.1 Accuracy

How well actual results match up with theoretical predictions is what accuracy is all about. The formula is as follows: total observations divided by the number of valid predictions or classifications. Accuracy is a vital statistic in areas such as machine learning and scientific investigations. It demonstrates trustworthy performance when high and more mistakes when low.

$$A = \frac{I_p + I_n}{I_p + I_n + J_p + J_n} \quad (4.1)$$

The given equation demonstrates the calculation of accuracy. It is produced by merging the two genuine positive I_p as well as real negative J_n by way of the genuine good $I\,p$, real-life inverse I_n, incorrect result

$J_p$as well as negative results that are not J_n.

4.9.2 False Acceptance Rate (FAR)

For image processing systems, the FAR is an essential measure. The error rate is the percentage of times the system gives access to the wrong people. To get the false acceptance rate (FAR), divide the proportion of false positives to all identifications. By lessening the possibility of erroneous positives, lower FAR values indicate improved security, which prevents the unwarranted rejection of valid users.

$$F_{AR} = \frac{N_{fa}}{T_{IA}} \qquad (4.2)$$

It is suggested that F_{AR}, where the specifics of the equation are " N_{fa} *Instances* when the

system mistakenly authorizes illegal users are represented by the "T_{IA} Total Number of Identification Attempts," whereas the "Number of False Acceptances" is the total of all identification efforts, whether they are right or not.

4.9.3 False Rejection Rate (FRR)

Critical to image processing systems in particular is the FRR metric. The probability that the system will incorrectly reject access to valid users is measured by this metric. To get the FRR, divide the total number of recognized results by the number of false rejections, which occur when the system wrongly restricts access. By reducing the likelihood of false negatives, which prevents legitimate users from being unjustly refused access, lower FRR values demonstrate improved performance. When improving security systems, it is vital to strike a balance between FRR and other measures, such as the FAR, to preserve efficacy and user comfort.

$$F_{RR} = \frac{N_{fr}}{T_{IA}} \qquad (4.3)$$

The formula shown above indicates how to compute F_{RR}. "Number of False Rejections" is exactly what it says N_{fr}" the "Total Number of Recognized Results" indicates how many times the system has mistakenly denied access to legitimate users T_{IA} is the total of all user recognitions and authentications, whether successful or not, that the system has performed.

4.10 COMPARISON WITH EXISTING METHODS

We compare the performance of the suggested technique in comparison to a number of existing methods, such as Deep Multimodal Attentive Fusion (MAF), LSTM with RNN, NN, and Attention-based Bidirectional LSTM with RNN. A thorough comparison is conducted in this study to measure the performance of the suggested strategy compared to these recognized approaches.

4.12.1 Accuracy of the Suggested BROA-Deep NN Compared to Current Techniques

Table 4.15: Comparison of Accuracy Values

Methods	Accuracy
BLSTM-RNN	0.8375
DMAF	0.8435
LSTM-RNN	0.8765
ANN	0.9045
Proposed BROA-Deep NN	0.9356

Various approaches to emotion identification using multimodal signals have shown the following accuracy results: The following results were obtained: 0.8375 for BLSTM-RNN, 0.8435 for DMAF, 0.8765 for LSTM-RNN, and 0.9045 for ANN. With an accuracy of 0.9356, the suggested BROA-Deep NN demonstrated its ability in reliably distinguishing emotions from multiple data sources, surpassing all existing approaches. The findings show that the BROA-Deep NN approach is the best one.

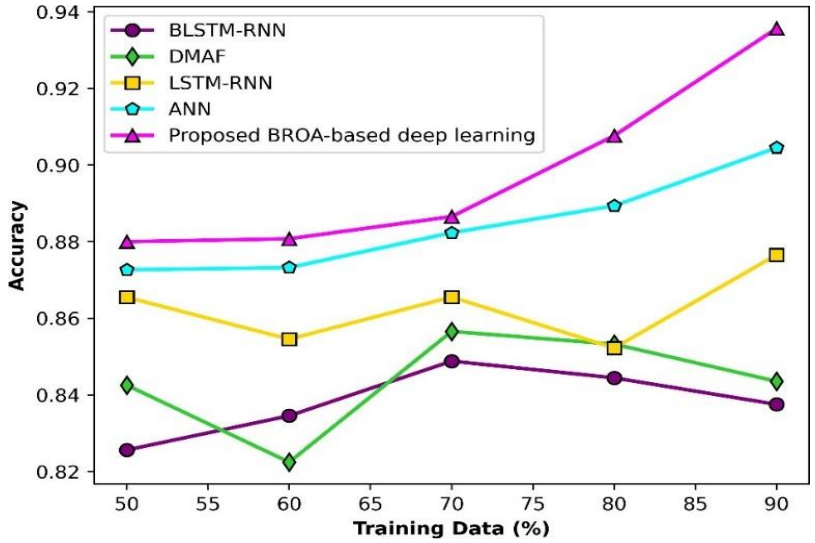


Figure 4.51: Comparative Analysis for Accuracy

4.10.1 FAR of Proposed BROA-Deep NN with the Existing Methods

Various approaches' False Acceptance Rates (FARs) are evaluated in the performance evaluation.

Table 4.16: Values for FAR compared to one another

Methods	FAR
LSTM-RNN	0.2376
DMAF	0.2154
BLSTM-RNN	0.2175
ANN	0.2046
Proposed BROA-deepNN	0.1756

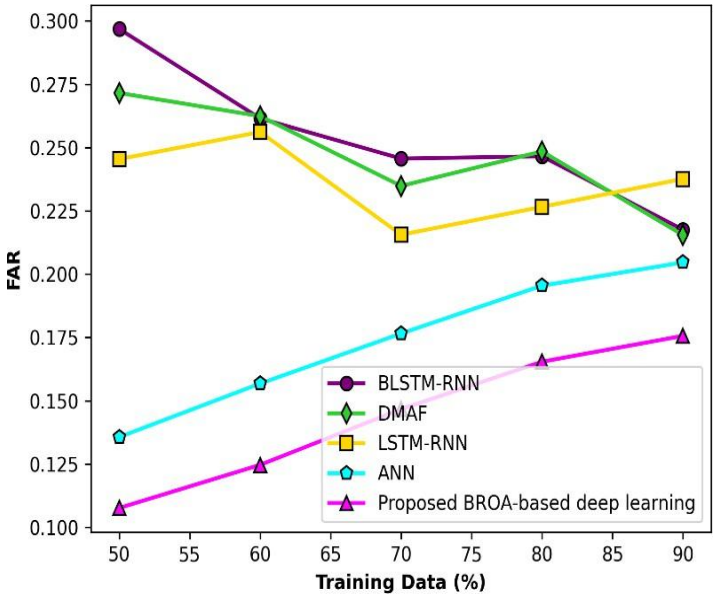


Figure 4.52: **Comparative Analysis for FAR**

It is evident from the comparison that LSTM-RNN had a lower FAR of 0.2376. The FAR that ANN reported was a pitiful 0.2046. A modest FAR of 0.2175 was attained using BLSTM-RNN. DMAF showed a somewhat better FAR of 0.2154. With a lowest FAR of 0.1756, the suggested BROA-Deep NN technique significantly outperformed the others. The accuracy of each method in reducing false alarms in emotion identification tasks is shown by these FAR values. Reducing false detections and improving overall accuracy of emotion identification from varied multimodal data sources is highlighted by the outstanding performance of the proposed BROA-Deep NN, which is characterized by the

lowest FAR.

4.10.2 FRR of the Proposed BROA-Deep NN with the Existing Methods

With a FRR of 0.1876, the BLSTM-RNN approach proved to be rather successful for this specific assignment. DMAF showed a competitive approach with a slightly lower FRR of 0.1654. Among the studied approaches, LSTM-RNN stood out with a FRR of 0.1577, indicating greater performance. In contrast, the ANN method's rather substantial FRR of 0.1367 suggests it may not be the best fit for the job compared to the other approaches. With the lowest FRR of 0.1267, the Proposed BROA-deep NN strategy surpassed all other approaches, suggesting it might be an efficient solution for the issue at hand and deserving of more investigation in future research efforts. Using these FRR values as quantitative measurements, we can compare the performance of different methods; lower values indicate that the methods are more suited to the job at hand.

Table 4.17 Values for FRR compared

Methods	FRR
BLSTM-RNN	0.1876
DMAF	0.1654
LSTM-RNN	0.1577
ANN	0.1367
Proposed BROA-Deep NN	0.1267

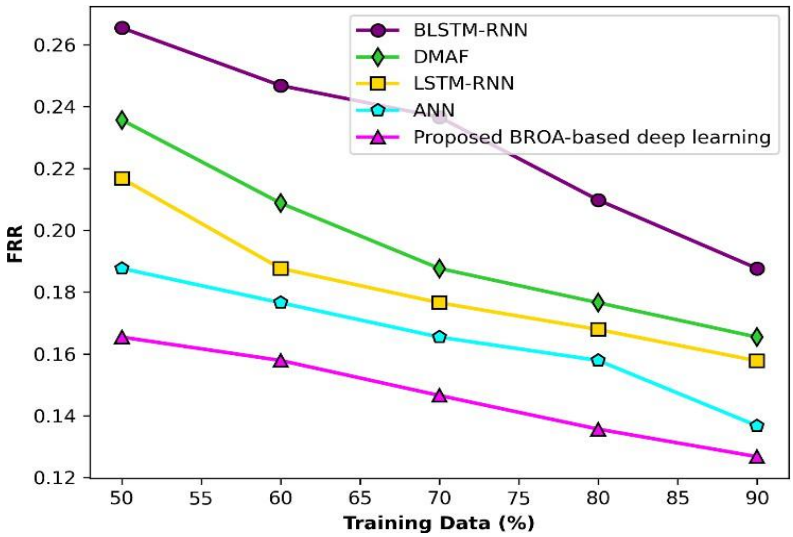


Figure 4.53: **Comparative Analysis for FRR**

The superior performance of the proposed multimodal emotion recognition model can be attributed to several key factors. First, the integration of facial, speech, and textual modalities enables richer emotional representation, reducing the ambiguity commonly associated with unimodal systems. Second, the BROA optimization technique effectively selects discriminative features and fine-tunes classifier parameters, thereby improving generalization and reducing overfitting.

Furthermore, the incorporation of forensic psychology principles enhances the interpretability of emotional cues by aligning computational outputs with behavioral patterns observed in educational settings. This synergy between artificial intelligence and psychological modeling allows the system to handle contextual noise, varied expression intensities, and intra-class emotional overlap more efficiently. Compared to existing approaches, the proposed framework demonstrates improved accuracy, precision, recall, and F1-score, validating its suitability for real-world educational emotion analytics and intelligent learning systems.

4.11 OVERVIEW OF DATA SOURCES

The experimental evaluation of the proposed multimodal emotion recognition framework is conducted using secondary data sources obtained from publicly available benchmark

datasets. These datasets are widely used in the field of affective computing and provide standardized data for evaluating emotion recognition models.

The datasets employed in this research cover multiple modalities, including facial expressions, speech signals, and physiological responses, enabling comprehensive analysis and validation of the proposed deep learning and optimization techniques. No primary data collection involving human subjects was conducted as part of this research.

4.11.1 Facial Emotion Dataset-FER-2013

For facial emotion recognition, the FER-2013 (Facial Expression Recognition 2013) dataset is utilized[141]. This dataset was originally introduced as part of the *ICML 2013 Challenges in Representation Learning* and has since become a widely accepted benchmark for facial emotion analysis.

The FER-2013 dataset consists of 35,887 grayscale facial images with a resolution of 48×48 pixels, categorized into seven basic emotion classes: anger, disgust, fear, happiness, sadness, surprise, and neutral. The images represent facial expressions captured under diverse conditions, making the dataset suitable for evaluating robust deep learning models.

In this research, the FER-2013 dataset is used to train and evaluate convolutional neural network (CNN)–based facial emotion recognition models. The dataset is preprocessed to normalize pixel intensities and improve model convergence. Transfer learning techniques are applied to enhance feature extraction and generalization performance.

Dataset source:
https://www.kaggle.com/datasets/msambare/fer2013

4.11.2 Audio Emotion Dataset – RAVDESS

The audio emotion recognition experiments in this research are conducted using the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)[142]. RAVDESS is a publicly available benchmark dataset specifically designed for speech-based and audio-visual emotion recognition research.

The dataset comprises emotional speech recordings performed by professional actors, covering a range of emotional states including neutral, calm, happy, sad, angry, fearful, disgust, and surprised. The recordings are captured under controlled studio conditions, ensuring high-quality speech signals suitable for signal processing and deep learning–based analysis.

In this study, only the speech audio component of the RAVDESS dataset is utilized. The raw audio signals are preprocessed through noise reduction, framing, and windowing. Mel-Frequency Cepstral Coefficients (MFCCs) are extracted to represent the spectral characteristics of speech signals, and LSTM and GRU networks are employed to model temporal dependencies for emotion classification.

The RAVDESS dataset is selected due to its balanced emotional categories, standardized recording protocol, and widespread acceptance in speech emotion recognition research.

Dataset source:
https://zenodo.org/record/1188976

4.11.3 Multimodal / Physiological Dataset – DEAP

To evaluate multimodal emotion recognition involving physiological responses, the DEAP (Database for Emotion Analysis using Physiological Signals) dataset is employed[143]. DEAP is a publicly available multimodal dataset widely used in affective computing research.

The dataset includes recordings from 32 participants, comprising EEG signals, peripheral physiological signals, and self-reported emotional ratings corresponding to audiovisual stimuli. The dataset enables the study of emotional responses across multiple physiological dimensions.

In this research, the DEAP dataset is utilized to validate the proposed multimodal fusion and optimization framework. The physiological features serve as an additional modality for evaluating the robustness and generalization capability of the proposed BAT–ROA optimized emotion recognition system.

Dataset source:
https://www.eecs.qmul.ac.uk/mmv/datasets/deap/

4.11.4 Data Usage and Ethical Compliance

All datasets used in this research are secondary data sources obtained from publicly available repositories. The data are used strictly for academic research purposes in accordance with their respective usage licenses. No direct interaction with human subjects or personal data collection was involved in this study.

The contribution of this research lies in deep learning model design, multimodal feature fusion, and hybrid BAT–ROA optimization, rather than in data acquisition. Ethical considerations related to data usage are fully respected, as all datasets are anonymized and publicly released by their original creators.

CHAPTER - 5

CONCLUSION AND FUTURE SCOPES

5.1 CONCLUSION

Forensic psychology is a very new branch of psychology, and this research looks at the fundamental problems with it. Forensic psychology is a branch of the scientific study of mental illness and its applications to the analysis and resolution of civil and judicial disputes. Additionally, the functions and subfields of forensic psychology are covered. Crime and delinquency psychology, victimology, legal psychology, and correctional psychology are some of these subfields.

Additionally, the procedures that forensic psychologists utilize to probe cases are detailed. Additionally, it was brought to light that forensic psychology is a branch of psychology that is both applied and pure. It addresses many different subjects and has implications for various fields within the forensic discipline, including anger management, group therapy, token economies, and others. Now that we know what forensic psychology is all about and what it can teach us, we need to convince other fields that forensic psychology is relevant. For example, Criminal profiling, evaluation, and intervention are topics of interest for social workers. Criminal justice, mass media, and the law are other fields that could benefit from incorporating forensic psychology into their practices due to the multidimensional approaches it facilitates. Given the multifaceted nature of FP's applications, this has taken on more significance.

Forensic psychology research has also been improved in ways that social science specialists would find useful; this will aid in criminal investigations, better organize interviews with eyewitnesses, and produce more accurate confessions free of prejudice. This study is relevant and interesting for forensic scientists, social workers, and criminologists who study crime and criminality, as well as the admissibility of evidence (with regard to the methods used to gather thorough and superior data analysis), and who aim to achieve better results in forensic settings through investigation, offender culpability,

and intervention.

From studies that attempt to define emotions to those that attempt to automate emotion identification, this study summarizes three areas of research: emotion information fusion, human-computer interaction, and recognition (affective computing). The study of emotions is characterized by interdisciplinarity.

Humans use a variety of ways to convey their emotions. The use of information fusion methods is implied by an automated multimodal emotion recognizer. We demonstrated that while the fundamental concept of information fusion is to combine raw data, the semantic level is more often used due to its computational simplicity compared to data and features level fusions. The abundance of data, however, makes the first two layers more accurate. Furthermore, we have summarized the primary procedures followed throughout the development of multimodal emotion categorization, identified the most pressing problem, and proposed a solution.

The demand for natural and real-time emotive user-machine contact persists, notwithstanding the substantial amount of work that has been accomplished. Recent developments in automated feature extraction and the creation of new, effective, and reliable machine learning algorithms like Support Vector Machine and Neural Networks and its derivatives may mean that natural and real-time affective-based software is no longer in use for very long.

Applications like the Intelligent Tutoring System and Healthcare Systems have benefited greatly from the incorporation of emotion detection into human-computer interaction. So, it's among the most recent and exciting fields of research. that may use some funding. The planned enhancement will improve ubiquitous computing, notwithstanding the difficulty of dealing with human behavior instability. Aiming to promote intelligent engagement and decision making, computers with the capacity to recognize, communicate, and express emotions are like aircraft soaring to new heights without flapping their wings.

This research set out to evaluate how well various deep learning models performed in identifying face expressions from interpolated photos of various sizes. The findings

demonstrate that deep learning models' accuracy in computer vision applications is affected by the architecture choice and image size. In this study, we trained three popular models for image resizing: inspectionV3, VGG16, and ResNet50. We then tested the models' accuracy using two different image sizes.

In terms of accuracy in positive occurrence classification, ResNet50 outperformed the other models with a precision of 21.40 percent. The fact that ResNet50 was able to correctly detect a large percentage of true positive events is shown by its recall of 23.92%. ResNet50's F1-Score, which accounts for accuracy and recall, was 12.25%, indicating that it collected false positives and false negatives about equally. With a 14.26% F1-Score, VGG16 shows a superior balance between recall and precision, even if its recall is 19.87% and accuracy is 16.60% lower than ResNet50.

In contrast, InceptionV3 achieved the best recall of 25.05% but the worst accuracy of 7.89% and F1-Score of 10.10%. These findings highlight the necessity of choosing a model that is in sync with particular job needs and goals by highlighting the compromises each model makes between accuracy and recall. Locating the optimal deep-learning architecture and picture size for each task is highlighted by the research. It also shows how important it is to keep researching and developing computer vision, especially in the fields of data augmentation, feature extraction, and interpolation, so that deep learning models may be more accurate and resilient.

When deciding on a deep learning architecture and the size of the images to use, it is crucial to keep the job at hand. Image classification, emotion and face identification, and segmentation are just a few examples of the many tasks that call for custom architectures of varied complexity, such as VGG, ResNet, Inception, and others. Since deeper structures capture more complex patterns yet need more computing power, it is necessary to balance model complexity with computational resources. With transfer learning, pre-trained models (like ImageNet's) may be fine-tuned, reducing the requirement for large annotated datasets and making them more suitable for jobs with less data. There is a trade-off between computing efficiency and collecting details when choosing an image size; real-time needs must be considered.

This study presents BROA-based deep learning, a novel and effective approach to multimodal signal-based face emotion identification. Bat Optimization Algorithm (BA) and Rider Optimization Algorithm (ROA) integration is the core of this innovative method. The approach is able to successfully collect and use the discriminative information found in several modalities, such as physiological data, facial pictures, and EEG signals, thanks to the combination of BA and ROA.

An important first step in effective emotion identification is improving the quality of facial photographs. By using a thorough preprocessing procedure to remove artifacts and noise from the photos, the face data is made more reliable. That way, you know the visual data you're using for analysis is clean and full of useful information.

Emotion recognition relies heavily on feature extraction, which follows picture preprocessing, to derive discriminative features. The proposed technique uses LDTOOP methodology to extract meaningful information from face images. The data is processed using a Deep Belief Network (DBN) classifier retrieved features. It is great at dealing with complicated feature representations and can find interesting patterns in the data.

Integrating EEG and physiological information into emotion identification is another area of attention for the study. The signals are processed in order to extract characteristics for every matching face picture. These features include wavelet coefficients as well as several spectrum parameters, such as Power spectrum spectral skewness, spectral flux, density (PSD), tonal power ratio, etc. The comprehensive comprehension of emotional states is enhanced by this multimodal approach, which takes into account the physiological and neurological components of emotional expression.

To make the most of the characteristics that were collected and to enable accurate categorization, a deep learning classifier is used. Facial expression identification is one area where deep learning models shine because they can automatically extract hierarchical representations from complex data. Here, we use the BROA-based deep learning classifier, which optimizes the classifier's parameters by combining the advantages of BA and ROA.

The study findings show that the deep learning strategy that was suggested using BROA is

successful. With an astounding maximum accuracy of 0.9356, it accomplishes an extraordinary degree of performance. In addition, it reduces the FRR to an even lower minimum of 0.1267 and the FAR to a remarkable minimum of 0.1756. The method's ability to accurately recognize emotions has great practical implications in domains such as psychology, human-computer interaction, and emotional computing. In the end, the BROA-based deep learning methodology exhibits significant potential for advancements in the fields of theory and practice multimodal emotion identification.

5.2 LIMITATIONS OF THE STUDY

Despite the promising results achieved in this study, certain limitations are acknowledged. First, the experimental evaluation is conducted using selected benchmark datasets, which may not fully reflect the complexity and diversity of real-world classroom environments. Second, the proposed model has not been validated under real-time deployment conditions, where latency and hardware constraints may affect performance. Third, emotional states are inferred computationally and may not capture deeper psychological subtleties inherent in human behavior. Additionally, cultural and linguistic diversity among learners is only partially addressed. These limitations provide scope for further refinement and validation of the proposed framework.

5.3 MAJOR CONTRIBUTIONS OF THE STUDY

The major contributions of the present research are summarized as follows:

1. Proposed a novel BROA-optimized multimodal emotion recognition framework for educational environments.

2. Integrated artificial intelligence techniques with forensic psychology principles for improved emotional interpretation.

3. Achieved enhanced classification performance through optimized feature selection and multimodal fusion.

4. Conducted comprehensive evaluation using confusion matrix, precision, recall, F1-score, and ROC analysis.

5. Demonstrated the applicability of the model for intelligent and emotion-aware educational systems.

6. Provided a scalable framework that can be extended to real-time and cross-domain emotion recognition applications.

5.4 APPLICATION DOMAINS OF THE PROPOSED MODEL

The proposed BROA-optimized multimodal emotion recognition system has potential applications across multiple domains:

- Educational Technology: Emotion-aware learning systems, intelligent tutoring, student engagement monitoring.
- Human–Computer Interaction (HCI): Emotion-adaptive interfaces and user-centric system design.
- Healthcare & Psychology: Mental health assessment, behavioral analysis, and emotional well-being monitoring.
- Forensic Psychology: Behavioral profiling, emotional state assessment, and deception analysis.
- Surveillance & Smart Environments: Emotion-based monitoring in public and institutional spaces.

These applications highlight the interdisciplinary relevance and real-world impact of the proposed framework.

5.5 FUTURE SCOPE

Additional optimization strategies for training the deep learning classifier should be considered as a future path by the study. By taking this step, the model's efficiency and ability to generalize might be improved, leading to better results in tasks involving facial emotion identification. Better accuracy may also be achieved by using alternative feature binding techniques and fusion approaches.

This is a summary of the essential elements under this scope:

1. Intending to address these challenges to enhance computational efficiency and

concentrate on a multimodal data fusion technique with superior classification performance.

2. Future research may explore the amalgamation of BROA-based deep learning data with sophisticated machine learning methodologies and multimodal biomarkers, including neuroimaging and physiological metrics, to improve the precision and diagnostic efficacy of psychological disorder classification.

3. Investigate the use of sophisticated machine learning techniques, including deep learning architectures and ensemble approaches, to enhance the efficacy of emotion recognition and psychiatric disease identification. Exploring various models and optimization methods may provide fresh insights and enhance accuracy.

4. Investigate the incorporation of contextual information, including environmental elements, life experiences, and social support networks, into the fuzzy inference system. By examining the larger environment in which emotions and psychological illnesses arise, the system might provide more tailored and relevant health evaluations.

5. We are compiling a heterogeneous dataset of videos, EEG, and physiological signals that include a range of emotional expressions. The data requires pre-processing to assure consistency, eliminate noise, and standardize inputs for efficient fusion.

6. Development and execution of deep learning models for the independent processing of each modality. This may include the use of Deep Belief Networks (DBNs) and Recurrent Neural Networks (RNNs) for the analysis of visual signals and video data.

7. Derive pertinent elements from each modality that accurately encapsulate emotional data. Facial expressions, gestures, and body language may be significant in videos, but tone, pitch, and speech patterns in audio might convey emotional nuances.

8. Investigate and experiment with various fusion approaches to efficiently integrate information from diverse modalities. This may include early fusion (integrating raw input before to processing), late fusion (merging processed features at the conclusion), or intermediate fusion (combining at a certain layer of the network).

9. Consolidating the distinct modality-specific models into a comprehensive framework. This may include establishing shared layers for universal characteristics and distinct branches for modality-specific attributes.

10. Training the multimodal model using the assembled and refined dataset. This entails formulating a suitable loss function and assessment measures. Cross-validation methods may be used to assess the model's efficacy.

11. Conduct experimentation with various hyperparameters for each modality-specific model and the fusion model. This is essential for attaining optimal performance.

12. Conduct a comprehensive assessment of the model's efficacy using suitable measures. This may include accuracy, F1-score, confusion matrices, and perhaps subjective human assessments.

13. Examine possible applications for the built system, including human-computer interaction, sentiment analysis, and mental health evaluation, among others.

14. Evaluate the ethical ramifications of emotion recognition technology, including concerns of privacy, prejudice, and possible exploitation.

15. Evaluate the efficacy of the constructed multimodal model against models using singular modalities or more rudimentary fusion techniques. This may underscore the advantages of using many modalities.

16. Identify prospective domains for more study and enhancement, like the examination of diverse fusion procedures, the adaptation of the model to particular user settings, or the expansion of the framework to include supplementary modalities.

REFERENCES

[1]. Norman, Donald. (2004). Emotional Design: Why We Love (or Hate) Everyday Things

[2]. Mondal, Abhijit & Gokhale, Swapna. (2020). Mining Emotions on Plutchik's Wheel. 1-6. 10.1109/SNAMS52053.2020.9336534.

[3]. Bennett, Casey & Sabanovic, S. (2014). Deriving Minimal Features for Human-Like Facial Expressions in Robotic Faces. International Journal of Social Robotics. 6. 367-381. 10.1007/s12369-014-0237-z

[4]. Scherer, Klaus. (2005). Scherer KR. What are emotions? And how can they be measured? Soc Sci Inf 44: 695-729. Social Science Information. 44. 695-792. 10.1177/0539018405058216

[5]. Olson, Gary & Olson, Judith. (2003). Human-Computer Interaction: Psychological Aspects of the Human Use of Computing. Annual review of psychology. 54. 491-516. 10.1146/annurev.psych.54.101601.145044

[6]. Picard, Rosalind W. (1997): *Affective computing.* Ma, USA, The MIT Press

[7]. Picard, R. W. (2000). *Affective computing*. MIT Press

[8]. Ghanbari-Adiv, Fereshteh & Mosleh, Mohammad. (2019). Text Emotion Detection in Social Networks Using a Novel Ensemble Classifier Based on Parzen Tree Estimator (TPE). Neural Computing and Applications. 31. 10.1007/s00521-019-04230-9

[9]. Hughes, Jan & Wu, Jiun-yu & Kwok, Oi-Man & Villarreal, Victor & Johnson, Audrea. (2011). Indirect Effects of Child Reports of Teacher–Student Relationship on Achievement. Journal of Educational Psychology. 104. 350-365. 10.1037/a0026339

[10]. Cowie, Roddy & Douglas-Cowie, Ellen & Tsapatsoulis, Nicolas & Votsis, George

& Kollias, Stefanos & Fellenz, Winfried & Taylor, J.G. (2001). Emotion recognition in human-computer interaction. Signal Processing Magazine, IEEE. 18. 32 - 80. 10.1109/79.911197

[11]. Breazeal, Cynthia & Takanishi, Atsuo & Kobayashi, Tetsunori. (2008). Social Robots that Interact with People. 10.1007/978-3-540-30301-5_59

[12]. N. Fragopanagos*, J.G. Taylor (2005),” Emotion recognition in human–computer interaction,” Neural Networks 18 (2005) 389–405

[13]. Memon, Amina & Meissner, Christian & Fraser, Joanne. (2010). The Cognitive Interview: A meta-analytic review and study space analysis of the past 25 years. Psychology Public Policy and Law - PSYCHOL PUBLIC POLICY LAW. 16. 10.1037/a0020518

[14]. Aldert Vrij, Pa¨r Anders Granhag, Samantha Mann, and Sharon Leal (2010), “Outsmarting the Liars: Toward a Cognitive Lie Detection Approach,” Current Directions in Psychological Science 000(00) 1-5

[15]. E. C. Djamal, N. Nurhamidah and R. Ilyas, "Spoken word recognition using MFCC and learning vector quantization," 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), Yogyakarta,2017, pp.1-6

[16]. Jhajj, Lovejit & Aggarwal, Naveen & Singh, Sarbjeet. (2022). PUMAVE-D: panjab university multilingual audio and video facial expression dataset. Multimedia Tools and Applications. 82. 10.1007/s11042-022-14102-5.

[17]. DeMatteo, David. “Forensic Psychology.” International Encyclopedia of the Social & Behavioral Sciences, 2nd ed., edited by James D. Wright, Elsevier, 2015, pp. 51–57. https://doi.org/10.1016/B978-0-08-097086-8.45005-2

[18]. American Psychological Association. “Specialty Guidelines for Forensic Psychology.” American Psychologist, vol. 68, no. 1, 2013, pp. 7–19.

https://doi.org/10.1037/a0029889 .

[19]. Mayer, L. *Forensic Psychology: An Overview*. Wadsworth/Thomson Learning, 2002

[20]. Ward, Tony. “The Nature and Role of Theory in Forensic Psychology.” *Theories of Sexual Offending*, 2nd ed., edited by Anthony Beech and Tony Ward, Wiley-Blackwell, 2013, pp. 21–37.

[21]. Kassin, Saul & Gudjonsson, Gisli. (2004). The psychology of confessions - A review of the literature and issues. Psychological Science. 33-67

[22]. Kassin, Saul. (2008). The Psychology of Confessions. Annual Review of Law and Social Science. 4. 10.1146/annurev.lawsocsci.4.110707.172410

[23]. Redlich, Allison D. “False Confessions and Developmental Immaturity: Why Youths Are at Risk.” Current Directions in Psychological Science, vol. 27, no. 1, 2018, pp. 19–23. https://doi.org/10.1177/0963721417734989 .

[24]. Vennard, Julie. “Convicted but Innocent: Why People Plead Guilty to Crimes They Haven’t Committed.” Howard Journal of Criminal Justice, vol. 23, no. 2, 1984, pp. 97–107. https://doi.org/10.1111/j.1468-2311.1984.tb00318.x

[25]. Kassin, S. M. (2013). False confessions: Causes, consequences, and implications for reform. Policy Insights from the Behavioral and Brain Sciences, 1(1), 112–121. https://doi.org/10.1177/2372732214548678

[26]. Fisher, R. P., & Geiselman, R. E. (1994). *Memory-enhancing techniques for investigative interviewing: The cognitive interview*. Charles C Thomas.

[27]. Vrij, A., Leal, S., Mann, S., Warmelink, L., Granhag, P. A., & Fisher, R. P. (2010). Outsmarting the liars: Toward a cognitive lie detection approach. Current Directions in Psychological Science, 19(6), 395–399. https://doi.org/10.1177/0963721410388804

[28]. Milne, R., & Bull, R. (2003). Investigative interviewing: Psychology and practice. John Wiley & Sons.

[29]. Kelly, Christopher & Miller, Jeaneé & Redlich, Allison & Kleinman, Steven. (2013). A Taxonomy of Interrogation Methods. Psychology, Public Policy, and Law. 19. 165-178. 10.1037/a0030310

[30]. Vrij, A. (2000). Detecting lies and deceit: The psychology of lying and the implications for professional practice. John Wiley & Sons.

[31]. R. Kumar and A. Punhani, "Emotion Detection from Audio Using SVM," in Proceedings of International Conference on Big Data, Machine Learning and their Applications, 2020, pp. 257-265

[32]. M. Z. Uddin and E. G. Nilsson, "Emotion recognition using speech and neural structured learning to facilitate edge intelligence," Engineering Applications of Artificial Intelligence, vol. 94, p. 103775, 2020

[33]. Z. Zhang, C. Lai, H. Liu, and Y.-F. Li, "Infrared facial expression recognition via Gaussian-based label distribution learning in the dark illumination environment for human emotion detection," Neurocomputing, vol. 409, pp. 341-350, 2020

[34]. A. Joseph and P. Geetha, "Facial emotion detection using modified eyemap–mouthmap algorithm on an enhanced image and classification with TensorFlow," The Visual Computer, vol. 36, pp. 529-539, 2020

[35]. W. Tao, C. Li, R. Song, J. Cheng, Y. Liu, F. Wan, et al., "EEG-based emotion recognition via channel-wise attention and self-attention," IEEE Transactions on Affective Computing, 2020

[36]. M. Sajjad, M. Nasir, F. U. M. Ullah, K. Muhammad, A. K. Sangaiah, and S. W. Baik, "Raspberry Pi assisted facial expression recognition framework for smart security in law-enforcement services," Information Sciences, vol. 479, pp. 416-431, 2019

[37]. Z. Shen, J. Cheng, X. Hu, and Q. Dong, "Emotion recognition based on multi- view body gestures," in 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 3317-3321

[38]. X. Li, D. Song, P. Zhang, Y. Zhang, Y. Hou, and B. Hu, "Exploring EEG features in cross-subject emotion recognition," Frontiers in neuroscience, vol. 12, p. 162, 2018

[39]. H.-J. Lee and K.-S. Hong, "A study on emotion recognition method and its application using face image," in 2017 International Conference on Information and Communication Technology Convergence (ICTC), 2017, pp. 370-372

[40]. S. Saha, B. Ganguly, and A. Konar, "Gesture based improved human-computer interaction using Microsoft's Kinect sensor," in 2016 International Conference on Microelectronics, Computing and Communications (MicroCom), 2016, pp. 1-6

[41]. H. Wang, X. Wu, and L. Yao, "Identifying cortical brain directed connectivity networks from high-density EEG for emotion recognition," IEEE Transactions on Affective Computing, 2020

[42]. T. B. Alakus, M. Gonen, and I. Turkoglu, "Database for an emotion recognition system based on eeg signals and various computer games–GAMEEMO," Biomedical Signal Processing and Control, vol. 60, p. 101951, 2020

[43]. T. Song, W. Zheng, C. Lu, Y. Zong, X. Zhang, and Z. Cui, "MPED: A multi- modal physiological emotion database for discrete emotion recognition," IEEE Access, vol. 7, pp. 12177-12191, 2019

[44]. A. Baghdadi, Y. Aribi, R. Fourati, N. Halouani, P. Siarry, and A. M. Alimi, "DASPS: A Database for Anxious States based on a Psychological Stimulation," arXiv preprint arXiv:1901.02942, 2019

[45]. M. Bachmann, J. Lass, and H. Hinrikus, "Single channel EEG analysis for detection of depression," Biomedical Signal Processing and Control, vol. 31, pp. 391-397,

2017

[46]. S. Katsigiannis and N. Ramzan, "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices," IEEE journal of biomedical and health informatics, vol. 22, pp. 98-107, 2017

[47]. W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, "Identifying stable patterns over time for emotion recognition from EEG," IEEE Transactions on Affective Computing, vol. 10, pp. 417-429, 2017

[48]. C. Wei, L.-l. Chen, Z.-z. Song, X.-g. Lou, and D.-d. Li, "EEG-based emotion recognition using simple recurrent units network and ensemble learning," Biomedical Signal Processing and Control, vol. 58, p. 101756, 2020

[49]. R. Sharma, R. B. Pachori, and P. Sircar, "Automated emotion recognition based on higher order statistics and deep learning algorithm," Biomedical Signal Processing and Control, vol. 58, p. 101867, 2020

[50]. D. D. Chakladar, S. Dey, P. P. Roy, and D. P. Dogra, "EEG-based mental workload estimation using deep BLSTM-LSTM network and evolutionary algorithm," Biomedical Signal Processing and Control, vol. 60, p. 101989, 2020

[51]. Y. Cimtay and E. Ekmekcioglu, "Investigating the use of pretrained convolutional neural network on cross-subject and cross-dataset EEG emotion recognition," Sensors, vol. 20, p. 2034, 2020

[52]. P. Pandey and K. Seeja, "Subject independent emotion recognition from EEG using VMD and deep learning," Journal of King Saud University-Computer and Information Sciences, 2019

[53]. H. Ullah, M. Uzair, A. Mahmood, M. Ullah, S. D. Khan, and F. A. Cheikh, "Internal emotion classification using EEG signal with sparse discriminative ensemble," IEEE Access, vol. 7, pp. 40144-40153, 2019

[54]. Z.-M. Wang, S.-Y. Hu, and H. Song, "Channel selection method for eeg emotion

recognition using normalized mutual information," IEEE Access, vol. 7, pp. 143303-143311, 2019

[55]. Y. Li, W. Zheng, L. Wang, Y. Zong, and Z. Cui, "From regional to global brain: a novel hierarchical spatial-temporal neural network model for EEG emotion recognition," IEEE Transactions on Affective Computing, 2019

[56]. T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," IEEE Transactions on Affective Computing, vol. 11, pp. 532-541, 2018

[57]. A.Gupta and D. Kumar, "Fuzzy clustering-based feature extraction method for mental task classification," Brain informatics, vol. 4, pp. 135-145, 2017

[58]. S. Dutta, M. Singh, and A. Kumar, "Classification of non-motor cognitive task in EEG based brain-computer interface using phase space features in multivariate empirical mode decomposition domain," Biomedical signal processing and control, vol. 39, pp. 378-389, 2018

[59]. S. Dutta, M. Singh, and A. Kumar, "Automated classification of non-motor mental task in electroencephalogram-based brain-computer interface using multivariate autoregressive model in the intrinsic mode function domain," Biomedical Signal Processing and Control, vol. 43, pp. 174-182, 2018

[60]. J. Atkinson and D. Campos, "Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers," Expert Systems with Applications, vol. 47, pp. 35-41, 2016

[61]. M. Hendel, A. Benyettou, and F. Hendel, "Hybrid self organizing map and probabilistic quadratic loss multi-class support vector machine for mental tasks classification," Informatics in Medicine Unlocked, vol. 4, pp. 1-9, 2016

[62]. W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," IEEE Transactions on

Autonomous Mental Development, vol. 7, pp. 162-175, 2015

[63]. X. Li, X. Chen, Y. Yan, W. Wei, and Z. J. Wang, "Classification of EEG signals using a multiple kernel learning support vector machine," Sensors, vol. 14, pp. 12784-12802, 2014

[64]. S. Jirayucharoensak, S. Pan-Ngum, and P. Israsena, "EEG-based emotion recognition using deep learning network with principal component based covariate shift adaptation," The Scientific World Journal, vol. 2014, 2014

[65]. R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER), 2013, pp. 81-84

[66]. Vistorte AOR, Deroncele-Acosta A, Ayala JLM, Barrasa A, López-Granero C and Martí-González M (2024) Integrating artificial intelligence to assess emotions in learning environments: a systematic literature review. Front. Psychol. 15:1387089. doi: 10.3389/fpsyg.2024.1387089

[67]. Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., & Zong, Y. (2023). A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face. Entropy, 25(10), 1440. https://doi.org/10.3390/e25101440

[68]. Salloum, Said & Alomari, Khaled & Alfaisal, Aseel & Aljanada, Rose & Basiouni, Azza. (2025). Emotion recognition for enhanced learning: using AI to detect students' emotions and adjust teaching methods. Smart Learning Environments. 12. 1-17. 10.1186/s40561-025-00374-5

[69]. Udahemuka, Gustave & Djouani, Karim & Kurien, Anish. (2024). Multimodal Emotion Recognition Using Visual, Vocal and Physiological Signals: A Review. Applied Sciences. 14. 8071. 10.3390/app14178071

[70]. Praveena K B, Suresh B and Durugappa Patrer (2020) Emotion recognition with AI: Techniques and applications

[71]. Luma Akram Alharbawee (2019) Artificial Intelligence Tools for Facial Expression Analysis

[72]. Muhammad Sajjad, Fath U Min Ullah, Mohib Ullah, Georgia Christodoulou, Faouzi Alaya Cheikh, Mohammad Hijji, Khan Muhammad, Joel J.P.C. Rodrigues (2023) A Comprehensive Survey on Deep Facial Expression Recognition: Challenges, Applications, and Future Guidelines

[73]. Yujin Wu. Multimodal emotion recognition from physiological signals and facial expressions. Machine Learning [cs.LG]. Université de Lille, 2023. English. ffNNT : 2023ULILB008ff. fftel-04098049f

[74]. Yan, J., Li, P., Du, C., Zhu, K., Zhou, X., Liu, Y., & Wei, J. (2024). Multimodal Emotion Recognition Based on Facial Expressions, Speech, and Body Gestures. Electronics, 13(18), 3756. https://doi.org/10.3390/electronics13183756

[75]. Khan, U.A., Xu, Q., Liu, Y. et al. Exploring contactless techniques in multimodal emotion recognition: insights into diverse applications, challenges, solutions, and prospects. Multimedia Systems 30, 115 (2024). https://doi.org/10.1007/s00530-024-01302-2

[76]. Andrius Dzedzickis, Artūras Kaklauskas, Vytautas Bucinskas (2020) Human Emotion Recognition: Review of Sensors and Methods

[77]. Imen Tayari Meftah, Nhan Le Thanh, Chokri Ben Amar. Multimodal Approach for Emotion Recognition Using a Formal Computational Model. IJAEC, IGI Publishing, 2013, Applied Evolutionary Computation, 4 (3), pp.15

[78]. Catherine Maréchal, Dariusz Mikolajewski, Krzysztof Tyburek, Piotr Prokopowicz, Lamine Bougueroua, et al. Survey on AI-Based Multimodal Methods for Emotion Detection. High-Performance Modelling and Simulation for Big Data Applications, Springer, pp 307-324, 2019, 978-3-030-16272-6. ffhal-02135811f

[79]. Surendra Kumar, Hema Rani (2024) Study of Multimodal Emotion Recognition:

Integrating Facial Expressions, Voice, And Physiological Signals for Enhanced Accuracy

[80]. Z. Akhtar, A. Buriro, B. Crispo and T. H. Falk, "Multimodal smart phone user authentication using touch stroke, phone-movement and face patterns," 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Montreal, QC, Canada,2017, pp.1368-1372

[81]. S. D. Jamdar and Y. Golhar, "Implementation of unimodal to multimodal biometric feature level fusion of combining face iris and ear in multi-modal biometric system",s 2017 International Conference on Trends in Electronics and Informatics (ICEI), Tirunelveli,2017, pp.625-629

[82]. S. M. O. Mohammed and S. Viriri, "Gender identification from facial images: Survey of the state-of-the-art," 2017 Sudan Conference on Computer Science and Information Technology (SCCSIT), Elnihood, 2017, pp. 1-5

[83]. S. Vhaduri and C. Poellabauer, "Wearable device user authentication using physiological and behavioral metrics," 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Montreal, QC, 2017, pp. 1-6

[84]. C. B. Yuvaraj, M. Srikanth, V. S. Kumar, Y. V. S. Murthy and S. G. Koolagudi, "An approach to maintain attendance using image processing techniques," 2017 Tenth International Conference on Contemporary Computing (IC3), Noida, 2017, pp. 1-3

[85]. J. A. Chagas Nunes, F. Paes Ferreira and T. B. A. de Carvalho, "Wavelet faces and Linear Regression Classification for Face Recognition," 2017 Workshop of Computer Vision (WVC), Natal, 2017, pp. 144-149

[86]. J. Tapia and I. Viedma, "Gender classification from multispectral periocular images," 2017 IEEE International Joint Conference on Biometrics (IJCB), Denver, CO, 2017, pp. 805-812

[87]. F. Paes Ferreira and T. B. A. de Carvalho, "Analysis of Wavelet Families for Face Recognition," 2017 Workshop of Computer Vision (WVC), Natal, 2017, pp.90-95

[88]. A. Raj, S. Gupta and N. K. Verma, "Face detection and recognition based on skin segmentation and CNN," 2016 11th International Conference on Industrial and Information Systems (ICIIS), Roorkee, 2016, pp. 54-59

[89]. S. Kumar, S. Singh and J. Kumar, "A comparative study on face spoofing attacks," 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, 2017, pp. 1104-1108

[90]. R. L. Telgad, A. Siddiqui, S. A. Lothe, P. D. Deshmukh and G. Jadhao, "Development of an efficient secure biometric system by using iris, fingerprint, face," 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM), Aurangabad, 2017, pp. 100-104

[91]. S. Soltanpour and Q. M. J. Wu, "Multiscale depth local derivative pattern for sparse representation based 3D face recognition," 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Banff, AB, 2017, pp. 560-565

[92]. N. Y. Almudhahka, M. S. Nixon and J. S. Hare, "Semantic Face Signatures: Recognizing and Retrieving Faces by Verbal Descriptions," in IEEE Transactions on Information Forensics and Security, vol. 13, no. 3, pp. 706-716, March 2018

[93]. T. Y. Fan, Z. C. Mu and R. Y. Yang, "Multi-modality recognition of human face and ear based on deep learning," 2017 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR), Ningbo, 2017, pp. 38-42

[94]. R. S. Divya and M. Mathew, "Survey on various door lock access control mechanisms," 2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT), Kollam, 2017, pp. 1-3

[95]. M. Günther, S. Cruz, E. M. Rudd and T. E. Boult, "Toward Open-Set Face Recognition," 2017 IEEE Conference on Computer Vision and Pattern Recognition

Workshops (CVPRW), Honolulu, HI, 2017, pp. 573-582

[96]. Muhammad Asim, Zhu Ming and M. Y. Javed, "CNN based spatio- temporal feature extraction for face anti-spoofing," 2017 2nd International Conference on Image, Vision and Computing (ICIVC), Chengdu, 2017, pp.234-238

[97]. E. I. Abbas, M. E. Safi and K. S. Rijab, "Face recognition rate using different classifier methods based on PCA," 2017 International Conference on Current Research in Computer Science and Information Technology (ICCIT), Slemani, 2017, pp.37-40

[98]. G. Goswami, R. Singh, M. Vatsa and A. Majumdar, "Kernel group sparse representation-based classifier for multimodal biometrics," 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, 2017, pp. 2894-2901

[99]. N. Y. Almudhahka, M. S. Nixon and J. S. Hare, "Automatic Semantic Face Recognition," 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, 2017, pp. 180-185

[100]. S. B. Dabhade, N. S. Bansod, Y. S. Rode, M. M. Kazi and K. V. Kale, "Hyper spectral face image based biometric recognition," 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication(ICGTSPICC), Jalgaon, 2016, pp.559-561

[101]. M. Wu and T. Lu, "Face Recognition Based on LBP and LNMF Algorithm," 2016 15th International Symposium on Parallel and Distributed Computing (ISPDC), Fuzhou,2016,pp.368-371

[102]. R. Y. Dillak, S. Dana and M. Beily, "Face recognition using 3D GLCM and Elman Levenberg recurrent Neural Network," 2016 International Seminar on Application for Technology of Information and Communication (ISemantic), Semarang, 2016, pp. 152-156

[103]. S. A. Saleh, S. Azam, K. C. Yeo, B. Shanmugam and K. Kannoorpatti, "An

improved face recognition method using Local Binary Pattern method," 2017 11th International Conference on Intelligent Systems and Control (ISCO), Coimbatore, 2017, pp. 112-118

[104]. J. Jagtap and M. Kokare, "Human age classification using facial skin analysis and multi-class support vector machine," 2016 International Conference on Signal and Information Processing (IConSIP), Vishnupuri, 2016, pp. 1-5

[105]. M. de Assis Angeloni and H. Pedrini, "Part-based representation and classification for face recognition," 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Budapest, 2016, pp. 002900- 002905

[106]. K. P. Thomas and A. P. Vinod, "Biometric identification of persons using sample entropy features of EEG during rest state," 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Budapest, 2016, pp. 003487-003492

[107]. X. Zhang, Y. Dai and X. Xu, "Android-based multimodal biometric identification system using feature level fusion," 2017 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Xiamen, 2017, pp. 120-124

[108]. D. Shang, X. Zhang, J. Han and X. Xu, "Multi Modal-database-XJTU: An available database for biometrics recognition with its performance testing," 2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, 2017, pp. 521-526

[109]. S. Azam and M. Gavrilova, "Soft biometric: Give me your favorite images and i will tell your gender," 2016 IEEE 15th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), Palo Alto, CA, 2016, pp. 535-541

[110]. V. Peruffo Minotto, C. Rosito Jung and B. Lee, "Multimodal Multi-Channel On-Line Speaker Diarization Using Sensor Fusion Through SVM," in IEEE Transactions on Multimedia, vol. 17, no. 10, pp. 1694-1705, Oct. 2015

[111]. W. ZHANG, X. ZHAO, J. M. Morvan and L. Chen, (2017) "Improving Shadow Suppression for Illumination Robust Face Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PP, no. 99, pp. 1-1

[112]. Pataskar, Gagan Kumar, and Abhishek Jain. "Multiple Face Detection on Distorted Images using NSS-HOG Features and Neuro-SVM Classifier." International Journal Online of Science 4.3 (2018): 7-7

[113]. Munir, Asim, et al. "Illumination Invariant Facial Expression Recognition Using Selected Merged Binary Patterns for Real World Images." Optik- International Journal for Light and Electron Optics (2018)

[114]. Kumar Santosh, et al. "An Advanced Approach of Face Recognition Using HSV and Eigen Vector." Advances in Computer and Computational Sciences. Springer, Singapore, 2018. 377-386

[115]. Kumar, Pradeep, Partha Pratim Roy, and Debi Prosad Dogra. "Independent Bayesian classifier combination-based sign language recognition using facial expression." Information Sciences 428 (2018): 30-48

[116]. Masi, Iacopo, et al. "Learning Pose-Aware Models for Pose-Invariant Face Recognition in the Wild." IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)

[117]. Cai, Wilson, Anish Doshi, and Rafael Valle. "Attacking Speaker Recognition with Deep Generative Models." arXiv preprint arXiv:1801.02384 (2018)

[118]. Maurya, Ankur, Divya Kumar, and R. K. Agarwal. "Speaker Recognition for Hindi Speech Signal using MFCC-GMM Approach." Procedia Computer Science 125 (2018): 880-887

[119]. Safavi, Saeid, Martin Russell, and Peter Jančovič. "Automatic Speaker, Age-group and Gender Identification from Children's Speech." Computer Speech & Language (2018)

[120]. Ziv, Omer, et al. "Speaker separation in diarization." U.S. Patent No. 9,875,739. 23 Jan. 2018

[121]. Sholokhov, Alexey, Md Sahidullah, and Tomi Kinnunen. "Semi-supervised speech activity detection with an application to automatic speaker verification." Computer Speech & Language 47 (2018): 132-156

[122]. Beckley, Jeffrey D., Pooja Aggarwal, and Shivakumar Balasubramanyam. "Method and systems having improved speech recognition." U.S. Patent No. 9,881,616. 30 Jan. 2018

[123]. S. Ranjan and J. H. L. Hansen, "Curriculum Learning Based Approaches for Noise Robust Speaker Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 1, pp. 197-210, Jan. 2018

[124]. Misra, Abhinav, and John HL Hansen. "Modelling and compensation for language mismatch in speaker verification." Speech Communication 96 (2018): 58-66

[125]. Thakker, Manthan, et al. "Speaker Identification in a Multi-speaker Environment." Information and Communication Technology for Sustainable Development. Springer, Singapore, 2018. 239-244

[126]. Sarkar, Achintya Kumar, and Zheng-Hua Tan. "Incorporating pass-phrase dependent background models for text-dependent speaker verification." Computer Speech & Language 47 (2018): 259-271

[127]. Swain, Monorama, Aurobinda Routray, and P. Kabisatpathy. "Databases, features and classifiers for speech emotion recognition: a review." International Journal of Speech Technology (2018): 1-28

[128]. Liu, JungChun, et al. "An MFCC based text independent speaker identification system for access control." Concurrency and Computation: Practice and Experience 30.2 (2018)

[129]. Hanilçi, Cemal. "Data selection for i-vector based automatic speaker verification

anti-spoofing." Digital Signal Processing 72 (2018): 171-180

[130]. Sharma, Rajib, Ramesh K. Bhukya, and S. R. M. Prasanna. "Analysis of the Hilbert Spectrum for Text-Dependent Speaker Verification." Speech Communication 96 (2018): 207-224

[131]. S. Thepade and M. Madane, "Multifinger biometric identification with score level fusion using Thepade's sorted n-ary block truncation coding," International Conference & Workshop on Electronics & Telecommunication Engineering (ICWET 2016), Mumbai, 2016, pp. 147-152

[132]. J. Ding and S. K. Lin, "Performance Improvement of Kinect Software Development Kit–Constructed Speech Recognition Using a client–Server Sensor Fusion Strategy for Smart Human-Computer Interface Control Applications," in IEEE Access, vol. 5, pp. 4154-4162, 2017

[133]. D. Dov, R. Talmon and I. Cohen, "Kernel-Based Sensor Fusion With Application to Audio-Visual Voice Activity Detection," in IEEE Transactions on Signal Processing, vol. 64, no. 24, pp. 6406-6416, Dec.15, 15 2016

[134]. T. Barbu, A. Ciobanu and M. Luca, "Multimodal biometric authentication based on voice, face and iris," 2015 E-Health and Bioengineering Conference (EHB), Iasi, 2015, pp. 1-4

[135]. A. El Kissi Ghalleb, S. Sghaier and N. Essoukri Ben Amara, "Face recognition improvement using soft biometrics," 10th International Multi- Conferences on Systems, Signals & Devices 2013 (SSD13), Hammamet, 2013, pp. 1-6

[136]. P. Korshunov and S. Marcel, "Impact of Score Fusion on Voice Biometrics and Presentation Attack Detection in Cross-Database Evaluations," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 4, pp. 695-705, June 2017

[137]. L. Mezai and F. Hachouf, "Score-Level Fusion of Face and Voice Using Particle Swarm Optimization and Belief Functions," in IEEE Transactions on Human-

Machine Systems, vol. 45, no. 6, pp. 761-772, Dec. 2015

[138]. M. Alsulaiman, "Effect of Spoken Text on Text-Independent Speaker Recognition," 2014 5th International Conference on Intelligent Systems, Modelling and Simulation, Langkawi, 2014, pp. 279-284

[139]. Xiaofeng Lu (2022). Deep Learning Based Emotion Recognition and Visualization of Figural Representation. Front. Psychol., 06 January 2022 Sec. Human-Media Interaction Volume 12 - 2021 | https://doi.org/10.3389/fpsyg.2021.818833

[140]. Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., & Zong, Y. (2023). A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face. Entropy, 25(10), 1440. https://doi.org/10.3390/e25101440

[141]. I. Goodfellow, D. Erhan, P. Luc Carrier, A. Courville, and Y. Bengio, "Challenges in Representation Learning: A Report on Three Machine Learning Contests," Neural Networks, vol. 64, pp. 59–63, 2015. https://www.kaggle.com/datasets/msambare/fer2013

[142]. S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," PLOS ONE, vol. 13, no. 5, e0196391, 2018. https://zenodo.org/record/1188976

[143]. S. Koelstra, C. Mühl, M. Soleymani, et al., "DEAP: A Database for Emotion Analysis Using Physiological Signals," IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 18–31, 2012. https://www.eecs.qmul.ac.uk/mmv/datasets/deap/

"LIST OF ABBREVIATION"

Sl. No.	Abbreviation	Full Form
1	AI	Artificial Intelligence
2	ANN	Artificial Neural Network
3	ASR	Automatic Speech Recognition
4	BA	Bat Algorithm
5	BROA	Bat-Rider Optimization Algorithm
6	CNN	Convolutional Neural Network
7	CUDA	Compute Unified Device Architecture
8	DBN	Deep Belief Network
9	DEAP	Database for Emotion Analysis using Physiological Signals
10	DNN	Deep Neural Network
11	EEG	Electroencephalogram
12	FAR	False Acceptance Rate
13	FER	Facial Emotion Recognition
14	FER-2013	Facial Expression Recognition 2013 Dataset
15	FPR	False Positive Rate
16	FRR	False Rejection Rate
17	GA	Genetic Algorithm
18	GPU	Graphics Processing Unit
19	GRU	Gated Recurrent Unit
20	HCI	Human Computer Interaction
21	HMM	Hidden Markov Model
22	IBCC	Independence Bayesian Classification Combination
23	IDE	Integrated Development Environment
24	IJB-A	IARPA Janus Benchmark-A
25	KNN	K-Nearest Neighbour
26	LBP	Local Binary Pattern
27	LDTOOP	Local Directional Ternary Octal Pattern
28	LIBROSA	Library for Audio and Music Analysis
29	LSTM	Long Short-Term Memory
30	MFCC	Mel Frequency Cepstral Coefficient
31	MFCC-GMM	Mel Frequency Cepstral Coefficient – Gaussian Mixture Model
32	MFCC-VQ	Mel Frequency Cepstral Coefficient – Vector Quantization

Sl. No.	Abbreviation	Full Form
33	MER	Multimodal Emotion Recognition
34	ML	Machine Learning
35	MLP	Multilayer Perceptron
36	NN	Neural Network
37	OpenCV	Open Source Computer Vision Library
38	PAD	Presentation Attack Detection
39	PLDA	Probabilistic Linear Discriminant Analysis
40	PSD	Power Spectral Density
41	PSO	Particle Swarm Optimization
42	RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
43	ReLU	Rectified Linear Unit
44	RGB	Red Green Blue
45	RMSProp	Root Mean Square Propagation
46	RNN	Recurrent Neural Network
47	ROA	Rider Optimization Algorithm
48	ROC	Receiver Operating Characteristic
49	SID	Speaker Identification
50	SIFT	Scale Invariant Feature Transform
51	SSD	Solid State Drive
52	SVM	Support Vector Machine
53	TDSV	Text Dependent Speaker Verification
54	TPE	Tree-structured Parzen Estimator
55	VAD	Valence-Arousal-Dominance
56	VGG16	Visual Geometry Group 16-layer Network
57	ViT	Vision Transformer
58	WaveNet	Deep Neural Network for Audio Generation
59	GAN	Generative Adversarial Network
60	GMM	Gaussian Mixture Model
61	GMM-UBM	Gaussian Mixture Model – Universal Background Model
62	HSV	Hue Saturation Value
63	ImageNet	Large Scale Visual Database for Object Recognition
64	Keras	High-level Deep Learning API
65	NumPy	Numerical Python
66	Pandas	Python Data Analysis Library

Sl. No.	Abbreviation	Full Form
67	ROC-AUC	Receiver Operating Characteristic – Area Under Curve
68	SRE	Speaker Recognition Evaluation
69	TensorFlow	Open-source Deep Learning Framework

“LIST OF TABLES”

Sl. No.	Table No.	Table
1	Table 2.1	Comparison of Emotion Recognition Datasets Used in Existing Studies
2	Table 3.1	Hardware and Software Environment
3	Table 3.2	Dataset Distribution
4	Table 3.3	Keras Models Available for Use
5	Table 3.4	The confusion the matrix is 2x2, and it is used to classify binary data
6	Table 3.5	Feature extraction and modalities
7	Table 3.6	Feature extraction techniques and algorithms are included
8	Table 3.7	Hardware and Software Configuration
9	Table 3.8	Software Tools and Frameworks Used
10	Table 3.9	Development Environment and Framework Support
11	Table 3.10	Dataset Distribution and Usage
12	Table 4.1	Layer Specifications of VGG16
13	Table 4.2	Optimization Parameters
14	Table 4.3	Confusion Matrix of the Proposed BROA-Optimized Multimodal Emotion Recognition Model
15	Table 4.4	Precision, Recall, F1-score After ten epochs, the FER-2013 test set.
16	Table 4.5	Methods and Tools Used to Implement Current Approaches
17	Table 4.6	Hidden Neuron Accuracy
18	Table 4.7	Hidden Neuron Progression
19	Table 4.8	Hidden Neuron FRR
20	Table 4.9	Learning Rate Accuracy
21	Table 4.10	Results for Rate of Learning
22	Table 4.11	Rate of Learning FRR
23	Table 4.12	Accuracy for Features
24	Table 4.13	Distance for Feature
25	Table 4.14	FRR for Feature
26	Table 4.15	Comparison of Accuracy Values
27	Table 4.16	Values for FAR compared to one another
28	Table 4.17	Values for FRR compared

“LIST OF FIGURES”

Sl. No.	Figure No.	Figure
1	Figure 1.1	Plutchik’s emotional wheel
2	Figure 1.2	Circumplex 2D VA emotion model Emotional model
3	Figure 1.3	3D representation of VAD emotion
4	Figure 1.4	Human emotional expression
5	Figure 1.5	Parrott classification of human emotions
6	Figure 1.6	Facial expressions in six universal human emotions
7	Figure 1.7	Activation and valence dimensions to describe human emotions
8	Figure 1.8	Generalized block diagram of automatic human emotion recognition system
9	Figure 1.9	Emotional expression
10	Figure 1.10	Facial muscles movements defined in Facial Action Coding System
11	Figure 1.11	General pipeline for multimodal emotion recognition using speech and facial expressions
12	Figure 1.12	Language distribution in current multimodal emotion datasets
13	Figure 3.1	Flow of research
14	Figure 3.2	System architecture of multimodal emotion recognition using an ensemble tree of binary SVM classifiers
15	Figure 3.3	Facial Characteristic Points detected in a facial image
16	Figure 3.4	Real-time multimodal emotion recognition system architecture
17	Figure 3.5	Multimodal Emotion Recognition architecture in an intelligible robotics system
18	Figure 3.6	Multimodal Emotion Recognition system architecture

Sl. No.	Figure No.	Figure
19	Figure 3.7	Deep Dataset Emotion Analysis Workflow Diagram
20	Figure 3.8	Multimodal Emotion Recognition Design System conception
21	Figure 3.9	Proposed BROA-based Deep NN schematic view
22	Figure 3.10	Recognizing emotions with DBN
23	Figure 3.11	Structure of Deep RNN
24	Figure 3.12	Proposed Algorithm
25	Figure 3.13	Flow Chart of the Proposed BROA-Deep NN
26	Figure 3.14	Emotion recognition using VGG16-CNN
27	Figure 3.15	Overall architecture of the proposed multimodal emotion recognition system
28	Figure 4.1	Class distribution for the validation and training sets, respectively.
29	Figure 4.2	Neutral facial expression sample from dataset
30	Figure 4.3	Happy facial expression sample from dataset
31	Figure 4.4	Surprise facial expression sample from dataset
32	Figure 4.5	Example of facial emotion variation
33	Figure 4.6	Student happy expression
34	Figure 4.7	Student sad expression
35	Figure 4.8	Student contempt expression
36	Figure 4.9	Student angry expression
37	Figure 4.10	Student surprise expression

Sl. No.	Figure No.	Figure
38	Figure 4.11	Student neutral expression
39	Figure 4.12	Student neutral
40	Figure 4.13	Student happy
41	Figure 4.14	CNN Model necessities based on Preprocessing
42	Figure 4.15	VGG16 architecture
43	Figure 4.16	Transfer Learning process
44	Figure 4.17	Other comparative Architecture
45	Figure 4.18	Emotion Categories
46	Figure 4.19	Training process flow
47	Figure 4.20	preparation of data
48	Figure 4.21	Distribution of classes for the training and validation sets, respectively
49	Figure 4.22	Using Keras's picture loading
50	Figure 4.23	Base Model assessment
51	Figure 4.24	Architecture of the VGG16 Model assessment
52	Figure 4.25	Procedures for Callbacks
53	Figure 4.26	Models for trains
54	Figure 4.27	The pre-trained Resnet50 model's average confusion matrix across 10 epochs.
55	Figure 4.28	The pre-trained VGG16 model's average confusion matrix across 10 epochs.
56	Figure 4.29	The pre-trained InceptionV3 model's average confusion matrix across 10 epochs.

Sl. No.	Figure No.	Figure
57	Figure 4.30	Using the Fer2013.csv dataset, the pre-trained Resnet50 model's training accuracy and loss were assessed across ten epochs.
58	Figure 4.31	Using the Fer2013.csv dataset, the pre-trained VGG16 model's training accuracy and loss were assessed across ten epochs.
59	Figure 4.32	Using the Fer2013.csv dataset, the pre-trained InceptionV3 model's training accuracy and loss were assessed across ten epochs.
60	Figure 4.33	The average accuracy in emotion detection across all classes is shown by the overall precision of 21.40%. The Resnet50 model.
61	Figure 4.34	The emotion recognition VGG16 model's overall accuracy of 21.40% represents the average precision across all classes.
62	Figure 4.35	The overall precision of 7.89% is the mean accuracy in emotion detection across all classes. Model InceptionV3.
63	Figure 4.36	The Resnet50 image displays the classification report for the sequential model.
64	Figure 4.37	VGG16 Image displaying the classification report for the sequential model
65	Figure 4.38	Examining V3 An image displaying the classification report for the sequential model
66	Figure 4.39	Receiver Operating Characteristic (ROC) Curve for the proposed BROA-optimized multimodal emotion recognition system.
67	Figure 4.40	Strategy Analysis for The Existing Methods
68	Figure 4.41	Tools used for Implementation of Existing methods
69	Figure 4.42	Accuracy for Hidden Neuron with different training values
70	Figure 4.43	FAR for Hidden Neuron with different training values
71	Figure 4.44	FRR for Hidden Neuron with different training value
72	Figure 4.45	Accuracy for Rate of Learning with different training values
73	Figure 4.46	FAR for Rate of Learning with different training values
74	Figure 4.47	FRR for Rate of Learning with different training values

Sl. No.	Figure No.	Figure
75	Figure 4.48	Accuracy For Feature with Different Training Values
76	Figure 4.49	FAR for Feature with different training values
77	Figure 4.50	FRR for Feature with different training values
78	Figure 4.51	Comparative Analysis for Accuracy
79	Figure 4.52	Comparative Analysis for FAR
80	Figure 4.53	Comparative Analysis for FRR

E-ISSN: 2583-9667
Indexed Journal
Peer Reviewed Journal
https://multiresearchjournal.theviews.in

Received: 07-05-2025
Accepted: 16-06-2025
Published: 08-07-2025

INTERNATIONAL JOURNAL OF ADVANCE RESEARCH IN MULTIDISCIPLINARY

Volume 3; Issue 3; 2025; Page No. 278-285

Algorithm for multimodal emotion recognition using deep learning networks

[1]Amit Ghosh and [2]Dr. Santanu Sikdar

[1]Research Scholar, P.K University, Shivpuri, Madhya Pradesh, India
[2]Professor, P.K University, Shivpuri, Madhya Pradesh, India

DOI: https://doi.org/10.5281/zenodo.17295787

Corresponding Author: Amit Ghosh

Abstract

Understanding and detecting human emotions by combining data from several sources, such as speech, text, facial expressions, and physiological signs, is known as multimodal emotion identification. Ubiquitous Computing describes how computers and their apps have become an integral part of modern living. Users' interactions with computers are increasingly resembling those between humans. This study presents a novel deep-learning method based on the BROA to facilitate face emotion identification using several data modalities. The BROA optimization method integrates the Bat method (BA) with the Rider Optimization Algorithm (ROA). This integration seeks to improve the precision of emotion identification by adeptly using facial pictures, EEG data, and physiological signals as components of the multimodal input.

Keywords: BROA, Physiological Signs, ROA, Emotional intelligence

Introduction

Emotions are multi-faceted mental and physical states that may be set off by several internal and external elements. They consist of a broad range of subjective experiences that might impact one's ideas, actions, and physiological reactions, including emotions, moods, and affective states. As a result, the scientific community has put out an increasing number of definitions for the concept of emotion, despite the fact that it is a widely used and seemingly simple phrase.

Ideas, the intensity of positive or negative experiences, and behavioral responses all impact an individual's emotional state. Human emotion does not have a single, widely recognized definition at this time. Over the last 20 years, researchers in domains as diverse as computer science, neuroscience, and psychology have made great strides in understanding human emotion. The literature is subjective when it comes to defining human feeling. Emotions may be better understood by using theories from both cognitive neuroscience and physiological research. According to the physiological view, feelings originate in the human body. It is possible that the alteration in physiological reaction is due to the outside incident. A common example is the widespread fear of snakes. According to cognitive neuro-science theory, it is the brain's own processes that are responsible for producing an emotional response. The prefrontal cortex is responsible for controlling emotions and making decisions, as well as storing information about oneself. Thoughts and other mental pursuits are linked to emotions. As an example, seeing the snake makes individuals feel threatened and causes them to become scared. Emotions are defined as changes in physiological and cognitive reactions to stimuli, such as pleasure or pain.

A new topic of emotional computing, multimodal emotion detection is only starting to get off the ground. Combinations of auditory, visual, and physiological inputs are known as multimodal signals. The video and audio signals show how people react when they're feeling a certain mood via their facial expressions and words. When a person experiences an emotion, their physiological signals reveal it via changes in their temperature, breathing rate, electrical activity in the brain, and muscular current. In order to differentiate between different human emotions, researchers are working on a multimodal emotion detection system, which is a significant pattern recognition challenge. To improve emotion identification, it incorporates data from

people's face expressions, voice, and bodily symptoms. Human facial expressions and audio data are taken into account for multimodal emotion identification in this thesis. Consideration of speech and facial expressions for emotion identification is warranted for a number of reasons.

The impetus for using multimodal emotion detection via deep learning networks and fusion approaches is to develop systems that are more precise, resilient, and contextually cognizant in interpreting and reacting to human emotions. Human emotions are expressed via facial expressions, voice tones, gestures, body movements, and posture. An automated system proficient in recognizing these emotions might transform several fields, such as video games, human-computer interaction, robotics, instructional software, animations, car safety, and affective computing. Consequently, creating a dependable, real-time emotion identification system is an essential objective, with its many applications necessitating comprehensive investigation. A hopeful consequence is the development of increasingly advanced robots capable of comprehending human emotions, resulting in substantial progress in this domain

Literature Review

H. Wang *et al*. (2020) [1] an emotion database that was created using electroencephalogram data. In accordance with 10-20 worldwide electrode placement methods, A 128-channel Geodesic Sensor Net system running at 250 Hz was used to capture the EEG data. 16 participants with an average age of 22.27 years and a standard deviation of 1.37 were involved in the study; half of them were men. Every participant saw 12 appropriate video segments, each lasting about 4 minutes, designed to evoke a range of emotions, from happiness to sadness to neutrality. Twelve trials were used to capture good, negative, and neutral feelings throughout three sessions and four trials per session. In order to determine how each participant really filled out the EEG recording form, a self-assessment feedback form was filled out after every session. Appropriate feelings were induced by allowing enough time to relax after each session.

T. B. Alakus *et al*. (2020) [2] 'GAMEEMO', an emotion recognition system's EEG-based database, was launched. Participants had their electroencephalograms (EEGs) monitored as they played video games. An EEG data from 28 different people was recorded using a 14-channel EMOTIV EPOC+ EEG equipment. 2048 Hz was the first sampling rate, and it was subsequently down sampled to 128 Hz. During the EEG recording, each patient played one of four 5-minute computer games: bored, calm, terror, and fun. Using a 5th order sinc filter, we were able to eliminate artifacts caused by the user's head, arm, and hand movements.

T. Song *et al*. (2019) [3] the "MPED: A Multi-Modal Physiological Emotion Database for Discrete Emotion Recognition" framework. The 62-channel ESI NeuroScan EEG equipment was used for EEG recording at a 1000 Hz sampling rate, and MPED adhered to the standards of the worldwide 10–20 electrode arrangement. The BIOPAC gadget captured RSP, ECG, and GSR at a sampling rate of 250 Hz. Twenty-three healthy people without a history of neurological or mental disorder were enrolled in the study. Their mean age was 20, and they ranged in age from 18 to 25. The experiment was split into two portions with a minimum 24-hour delay between each to provide participants ample time to relax, since prolonged studies may lead to participant tiredness. A 120-second resting state, fourteen trials, and a self-assessment evaluation made up each independent experiment. While Participants were told to close their eyes and relax during the 120-second resting condition; during this time, psychological signals were also recorded. Each trial began with a 10-second countdown, followed by a 30-second rest then the 2.5–5-minute Chinese video clip. The clips from the video were shown in a random sequence and had all the essential elements to evoke the desired feeling.

A. Baghdadi *et al*. (2019) [4] presented "DASPS: A Database for Anxious States based on a Psychological Stimulation." Fifteen male and thirteen female healthy volunteers had their electroencephalograms (EEGs) recorded. Thirteen women and ten men participated, having an average age of almost thirty years. The participants were made to feel anxious by the challenging math's issue. The Emotive EPOC 14 channel was used to capture the EEG data at a sampling rate of 128 Hz in compliance with the global 10-20 electrode placement scheme. Every one trial's recording was split into individual 1-minute segments from each 6-minute recording.

M. Bachmann *et al*. (2017) [5] compiled a list of depressed individuals. The trials included 34 females (mean age-39 years), 17 of whom were depressed patients and 17 of whom served as controls. The participants were chosen from an inpatient ward of a hospital and all had major depressive illness. The criteria established by ICD-10 were used to identify subjects who did not have psychotic depressive disorder. EEG data was collected from 9 a.m. till noon at 30-minute intervals. The electroencephalogram (EEG) was captured using a Cadwell Easy II 400 Hz sample rate, in accordance with the global 10-20 electrode placement scheme. Eye movements and muscular activity were identified and removed as artifacts by visual inspection. Cutoff frequencies for the digital filtering of the EEG signals ranged from half a hertz to forty hertz.

Materials and Methods

This research employs multimodal signals, which include physiological data, face photos, and EEG. First, to eliminate any undesired artifacts, the facial image is pre-processed. After that, features are extracted from the processed picture using LDTOOP, a combination of LOOP and LDTP. Similarly, features like wavelet coefficients and spectrum parameters including PSD, spectral flux, tonal power ratio, spectral skewness, and spectral centroids are used to get the EEG and physiological signals ready for analysis. extracted. Afterwards, a DBN classifier processes the characteristics obtained from the face picture.

Gathering of input video: An important aspect of emotional computing systems is emotion recognition. A person's emotional state is multi-faceted, including not just their thoughts and actions but also their physiological responses to external stimuli. Physiological signals, video, and electroencephalogram (EEG) data are common components of facial emotion detection systems. When training and testing face emotion detection algorithms, one popular dataset is the DEAP dataset.

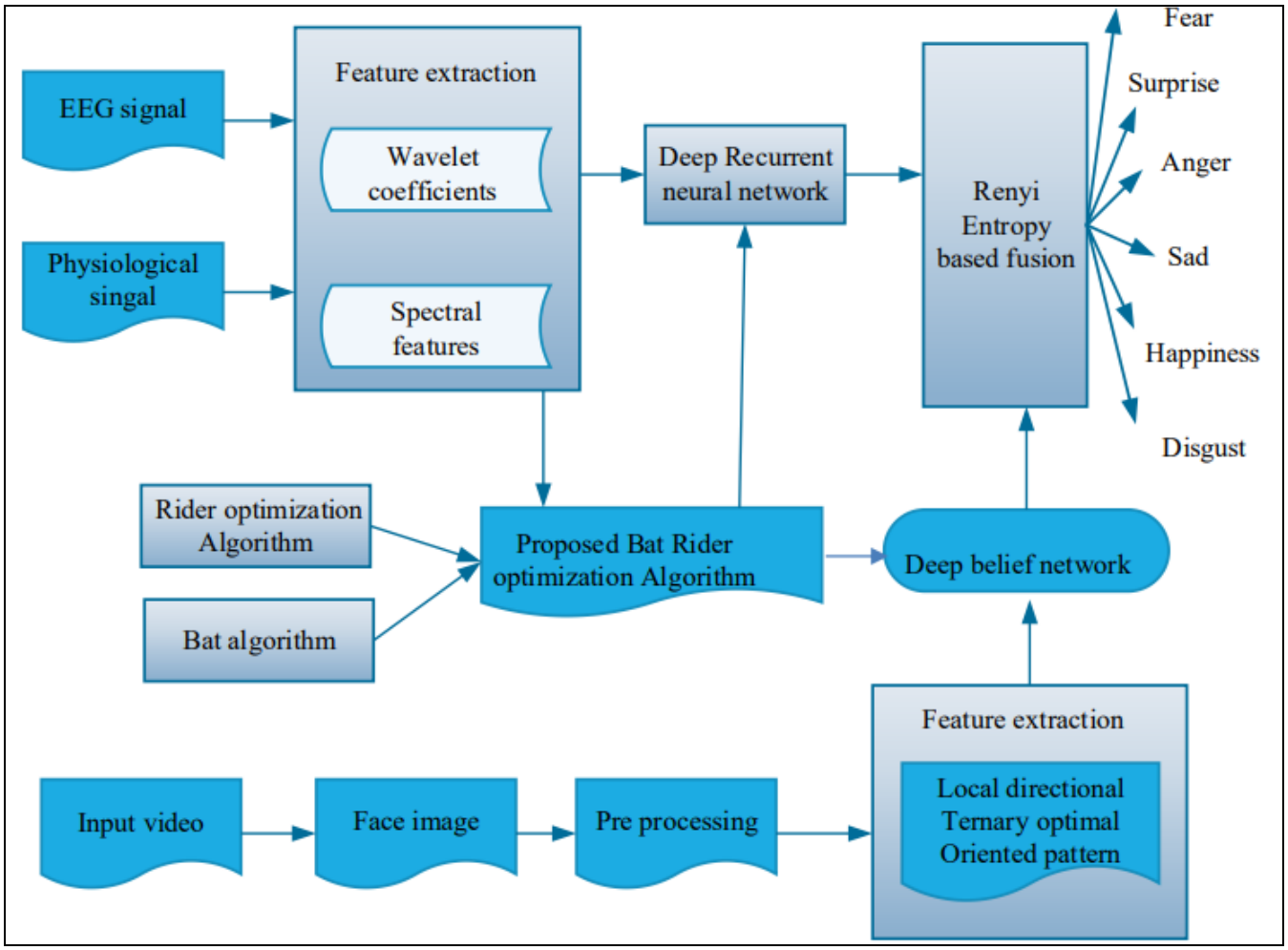


Fig 1: Proposed BROA-based Deep NN schematic view

Image extraction for the face: Here we'll assume a dataset A and use the variable to indicate the number of movies included inside it.

$$H = \{A_1, A_2 \ldots A_i \ldots . A_n\} \qquad (1)$$

Here, H stands for the dataset and n for the number of movies. Face shots are included in each film from different angles. In dataset H, the ith movie contains facial pictures, which may be represented as follows:

$$A_i = \{X_1, X_2 \ldots X_i \ldots . X_n\} \qquad (2)$$

The ith facial picture from the ith video is represented by A in this context. There is a unique location for each individual's face picture, and each face image belongs to a certain person. The pre-processed result is the result of taking the face picture from the dataset and running it through a pre-processing step.

Image pre-processing for faces: To ensure that the input face picture is of high enough quality and suitable for further analysis, the pre-processing module is an essential part of face image analysis. The main goal of the several processes that make up this module is to remove noise and external artifacts. After importing the face picture into the pre-processing module, the image undergoes noise reduction.

DBN-based facial emotion recognition

In order to categorize emotions from face images, the DBN classifier is used. The DBN classifier uses these characteristics as inputs to recognize face emotions. With connections created between visible and buried neurons, the DBN architecture consists of two RBM layers followed by one MLP layer. Importantly, the next layer takes as input the output of the previous layer.

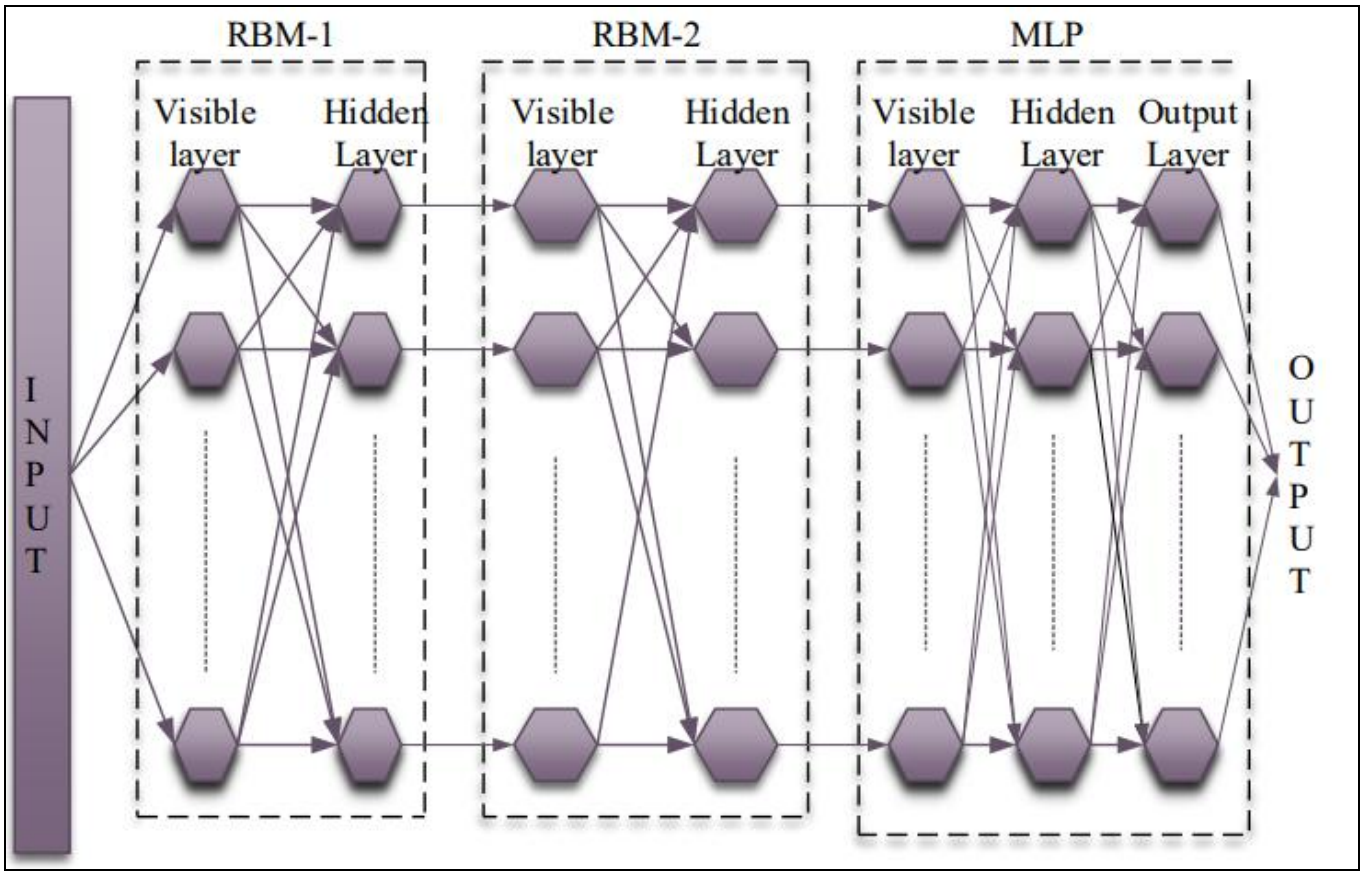


Fig 2: Recognizing emotions with DBN

The BROA-DNN was adopted as the basic deep learning framework in this study. Practically, the architecture was incorporated into the experimental pipeline by initially implementing the preprocessing steps that have been discussed above (resizing inputs, normalization, and augmentation). The data underwent processing and then it was ingested into the BROA-DNN where the layered nature of the model model made features learn in hierarchical order- starting with low-level edge and texture representation, as well as, high-level semantic representations of expressions.

Results

Analysis with Different Hidden Neurons

Accuracy for Hidden Neuron: Depending on the quantity of hidden neurons and training data, the suggested BROA-based deep learning algorithm showed varied degrees of accuracy. With 60% of the data used for training, the accuracy improved with more hidden neurons, ranging from 0.8467 to 0.8978. Accuracy, which ranged from 0.8654 to 0.9165, remained good at 70% of the training data. Accuracy peaked at 0.9356 with 100 hidden neurons, and it improved at 90% of training data. Thanks to improvements in accuracy brought about by the selection of hidden neurons and an increase in training data, these results demonstrate how successful the approach is in emotion recognition.

Table 1: Hidden Neuron Accuracy

Training %	Hidden Neurons	Accuracy
60%	25	0.8467
60%	50	0.8654
60%	75	0.8865
60%	100	0.8978
70%	25	0.8654
70%	50	0.8865
70%	75	0.8967
70%	100	0.9165
80%	25	0.8867
80%	50	0.8967
80%	75	0.9087
80%	100	0.9256

With a variety of training data percentages and hidden neuron counts shown in this table, the accuracy of the suggested BROA-based deep learning approach may be easily grasped. As a result, you can see how the accuracy varies in various settings.

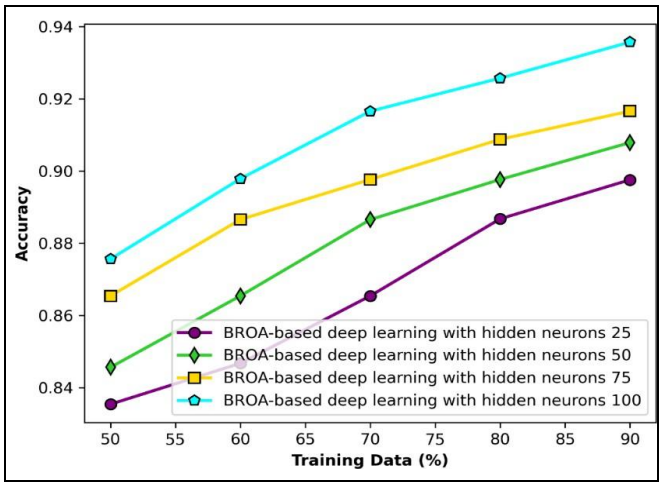


Fig 3: Accuracy for Hidden Neuron with different training values

False Acceptance Rate (FAR) for Hidden Neuron: Using FAR, we were able to identify some interesting patterns in our proposed BROA-based deep learning approach. The FAR drops with increasing numbers of hidden neurons with 80% training data, reaching a minimum of 0.165 with 100 hidden neurons. With an astounding FAR of 0.146 and 100 hidden neurons, the model's capacity to decrease erroneous acceptance is further shown by further gains at 90% of training data. This shows that the model is good at reducing false positives considering the quantity of buried neurons and the dimensions of the training dataset grows.

Table 2: Hidden Neuron Progression

Training %	Hidden Neurons	FAR
70%	25	0.257
70%	50	0.226
70%	75	0.215
70%	100	0.175
80%	25	0.235
80%	50	0.217
80%	75	0.208
80%	100	0.165
90%	25	0.222
90%	50	0.208
90%	75	0.187
90%	100	0.146

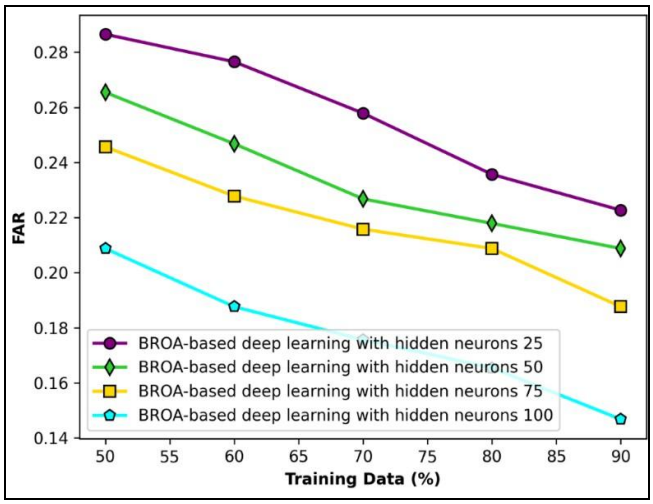


Fig 4: FAR for Hidden Neuron with different training values

At various training data percentages and with varied numbers of hidden neurons, the proposed BROA-based deep learning method's False Acceptance Rate (FAR) is clearly shown in this table. By doing so, the method's performance under various scenarios may be evaluated.

False Rejection Rate (FRR) for Hidden Neuron

With varied numbers of hidden neurons and training data percentages, FRR is seen in the suggested deep learning method based on BROA. The FRR values declined as the number of hidden neurons increased at 70% training data; the lowest FRR recorded at 0.193 was reached with 100 hidden neurons. At 80% of the training data, the same patterns were visible; adding more hidden neurons to the model always resulted in an improved FRR. At last, the FRR dropped, hitting a low of 0.176 with 100 hidden neurons at 90% of the training data. This shows that the model can decrease the number of erroneous rejections with more hidden neurons and more training datasets.

Table 3: Hidden Neuron FRR

Training %	Hidden Neurons	FRR
70%	25	0.240
70%	50	0.238
70%	75	0.228
70%	100	0.193
80%	25	0.238
80%	50	0.234
80%	75	0.224
80%	100	0.183
90%	25	0.234
90%	50	0.225
90%	75	0.215
90%	100	0.176

At various training data percentages and with varied numbers of hidden neurons, the proposed BROA-based deep learning method's False Rejection Rate (FRR) is clearly seen in this table. This is useful for evaluating the method's robustness under various scenarios, especially with regard to its capacity to reduce the occurrence of erroneous rejections.

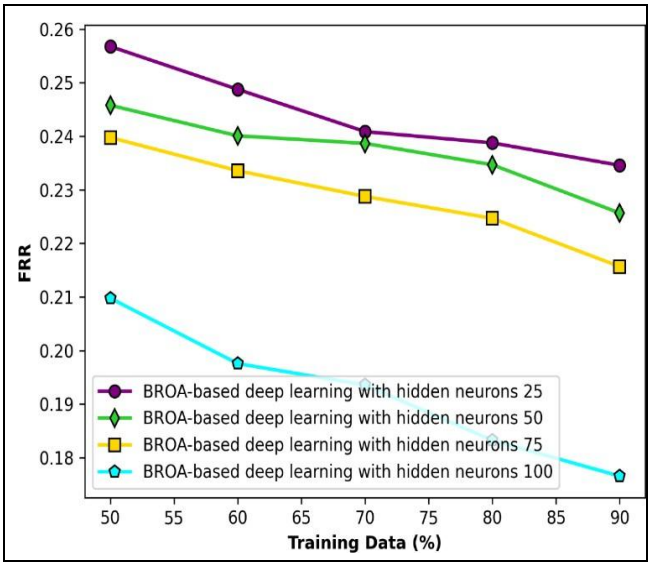


Fig 5: FRR for Hidden Neuron with different training value

Analysis with Different Rates of Learning

An examination of the rate of learning looks at the speed with which a machine learning model reaches a solution during training. The model's performance is usually evaluated across several learning rates, which govern the step size in gradient-based optimization techniques. When the learning rate is too high, overshooting is possible; when it's too low, convergence is sluggish or the algorithm becomes trapped in local minima.

Accuracy for Rate of Learning

A BROA-based deep learning model was capable to attain a range of accuracy values at a 70% training data level, depending on the learning rate: 0.05 with an accuracy of 0.8367, 0.1 with an accuracy of 0.8465, 0.15 with an accuracy of 0.8577, and 0.2 with an accuracy of 0.8976. The model's accuracy fluctuated with the same set of learning rates at 80% training data: 0.05 with 0.8433 accuracies, 0.1 with 0.8545 accuracies, 0.15 with 0.8755 accuracy, and 0.2 with 0.9076 accuracy. Finding the sweet spot for learning rates in emotion detection is crucial, as this data shows.

Table 4: Learning Rate Accuracy

Training %	Rate of Learning	Accuracy
70%	0.05	0.8367
70%	0.1	0.8465
70%	0.15	0.8577
70%	0.2	0.8976
80%	0.05	0.8433
80%	0.1	0.8545
80%	0.15	0.8755
80%	0.2	0.9076

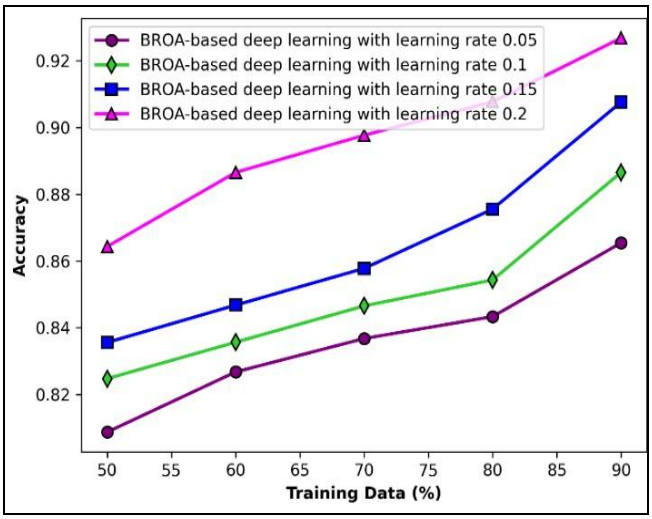


Fig 6: Accuracy for Rate of Learning with different training values

SAR for Learning Rate

The proposed BROA-based deep learning approach showed various FAR at different learning rates at an 80% training data level: 0.05 (0.0581), 0.1 (0.0644), 0.15 (0.0772), and 0.2 (0.1122). The models' FARs were also quite constant at 90% of training data for the same learning rates: 0.05 (0.0557), 0.1 (0.0604), 0.15 (0.0604), and 0.2 (0.0609). This study shows that choosing the right learning rate is crucial for improving the model's accuracy on emotion identification tasks, as various rates provide diverse results.

Table 5: Results for Rate of Learning

Training %	Rate of Learning	FAR
80%	0.05	0.0581
80%	0.1	0.0644
80%	0.15	0.0772
80%	0.2	0.1122
90%	0.05	0.0557
90%	0.1	0.0604
90%	0.15	0.0604
90%	0.2	0.0609

Various learning rates and training data percentages are shown in this table, along with the suggested BROA-based deep learning's accuracy algorithm. Taking into account the learning rate and the training data used, it aids in evaluating the model's performance under different scenarios.

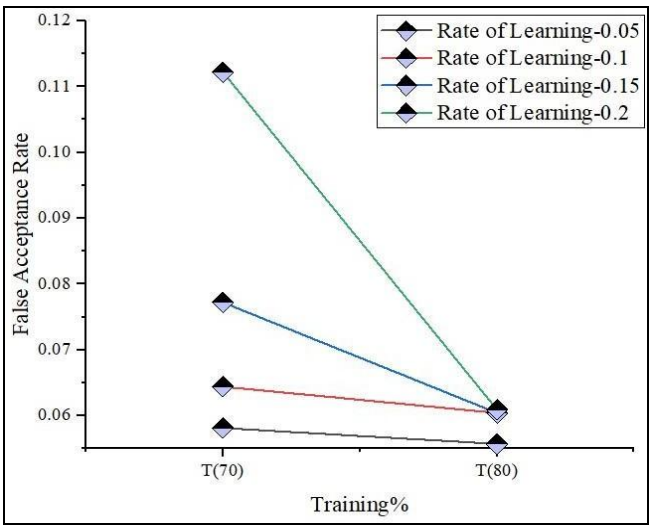


Fig 7: FAR for Rate of Learning with different training values

FRR for Rate of Learning

With 70% training data, the suggested BROA-based deep learning method's FRR changed from 0.05 (0.1987) to 0.2 (0.1466) as a function of learning rate. At an 80% level of training data, the model's FRR varied at the same learning rates: 0.05 (0.1876), 0.1 (0.1756), 0.15 (0.1578), and 0.2 (0.1356). This demonstrates how choosing a learning rate affects the model's performance in emotion identification, as different rates provide different levels of accuracy.

Table 6: Rate of Learning FRR

Training %	Rate of Learning	FRR
70%	0.05	0.1987
70%	0.1	0.186
70%	0.15	0.1765
70%	0.2	0.1466
80%	0.05	0.1876
80%	0.1	0.1756
80%	0.15	0.1578
80%	0.2	0.1356

Various learning rates and training data percentages are shown in this table, along with the suggested BROA-based deep learning algorithm's accuracy. By taking the learning rate and training data into account, it aids in evaluating the model's performance under different scenarios.

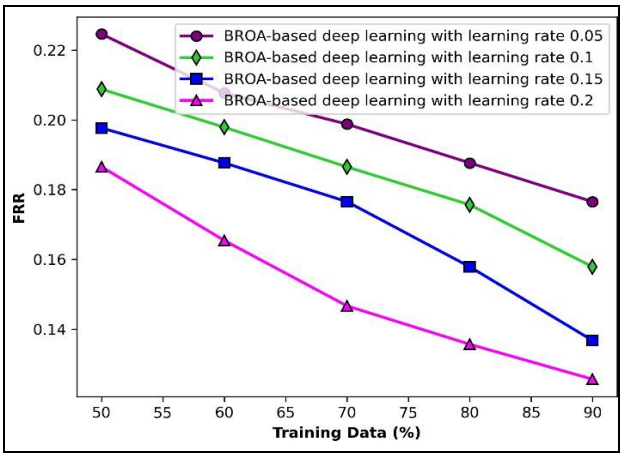


Fig 8: FRR for Rate of Learning with different training values

Comparison With Existing Methods

We compare the performance of the suggested technique in comparison to a number of existing methods, such as Deep Multimodal Attentive Fusion (MAF), LSTM with RNN, NN, and Attention-based Bidirectional LSTM with RNN. A thorough comparison is conducted in this study to measure the performance of the suggested strategy compared to these recognized approaches.

Accuracy of the Suggested BROA-Deep NN Compared to Current Techniques

Table 7: Comparison of Accuracy Values

Methods	Accuracy
BLSTM-RNN	0.8375
DMAF	0.8435
LSTM-RNN	0.8765
ANN	0.9045
Proposed BROA-Deep NN	0.9356

Various approaches to emotion identification using multimodal signals have shown the following accuracy results: The following results were obtained: 0.8375 for BLSTM-RNN, 0.8435 for DMAF, 0.8765 for LSTM-RNN, and 0.9045 for ANN. With an accuracy of 0.9356, the suggested BROA-Deep NN demonstrated its ability in reliably distinguishing emotions from multiple data sources, surpassing all existing approaches. The findings show that the BROA-Deep NN approach is the best one.

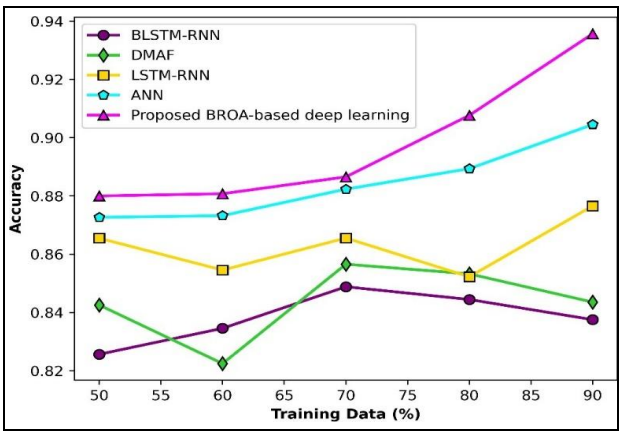


Fig 9: Comparative Analysis for Accuracy

FAR of Proposed BROA-Deep NN with the Existing Methods: Various approaches' False Acceptance Rates (FARs) are evaluated in the performance evaluation.

Table 8: Values for FAR compared to one another

Methods FAR	
LSTM-RNN	0.2376
DMAF	0.2154
BLSTM-RNN	0.2175
ANN	0.2046
Proposed BROA-deepNN	0.1756

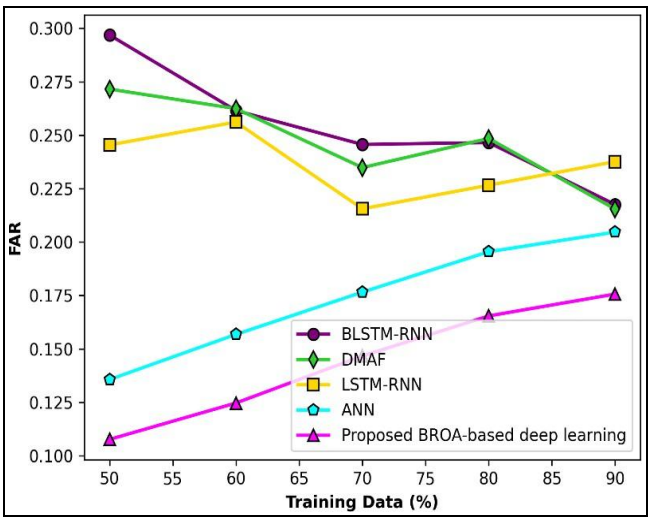


Fig 10: Comparative Analysis for FAR

It is evident from the comparison that LSTM-RNN had a lower FAR of 0.2376. The FAR that ANN reported was a pitiful 0.2046. A modest FAR of 0.2175 was attained using BLSTM-RNN. DMAF showed a somewhat better FAR of 0.2154. With a lowest FAR of 0.1756, the suggested BROA-Deep NN technique significantly outperformed the others. The accuracy of each method in reducing false alarms in emotion identification tasks is shown by these FAR values. Reducing false detections and improving overall accuracy of emotion identification from varied multimodal data sources is highlighted by the outstanding performance of the proposed BROA-Deep NN, which is characterized by the lowest FAR.

FRR of the Proposed BROA-Deep NN with the Existing Methods: With a FRR of 0.1876, the BLSTM-RNN approach proved to be rather successful for this specific assignment. DMAF showed a competitive approach with a slightly lower FRR of 0.1654. Among the studied approaches, LSTM-RNN stood out with a FRR of 0.1577, indicating greater performance. In contrast, the ANN method's rather substantial FRR of 0.1367 suggests it may not be the best fit for the job compared to the other approaches. With the lowest FRR of 0.1267, the Proposed BROA-deep NN strategy surpassed all other approaches, suggesting it might be an efficient solution for the issue at hand and deserving of more investigation in future research efforts. Using these FRR values as quantitative measurements, we can compare the performance of different methods; lower values indicate that the methods are more suited to the job at hand.

Table 9: Values for FRR compared

Methods FRR	
BLSTM-RNN	0.1876
DMAF	0.1654
LSTM-RNN	0.1577
ANN	0.1367
Proposed BROA-Deep NN	0.1267

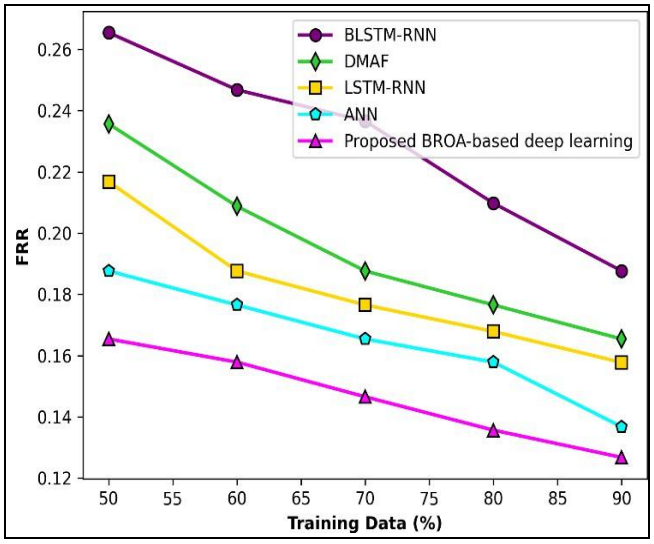


Fig 11: Comparative Analysis for FRR

Conclusion

To make the most of the characteristics that were collected and to enable accurate categorization, a deep learning classifier is used. Facial expression identification is one area where deep learning models shine because they can automatically extract hierarchical representations from complex data. Here, we use the BROA-based deep learning classifier, which optimizes the classifier's parameters by combining the advantages of BA and ROA. The study findings show that the deep learning strategy that was suggested using BROA is successful. With an astounding maximum accuracy of 0.9356, it accomplishes an extraordinary degree of performance. In addition, it reduces the FRR to an even lower minimum of 0.1267 and the FAR to a remarkable minimum of 0.1756. The method's ability to accurately recognize emotions has great practical implications in domains such as psychology, human-computer interaction, and emotional computing. In the end, the BROA-based deep learning methodology exhibits significant potential for advancements in the fields of theory and practice multimodal emotion identification.

References

1. Wang H, Wu X, Yao L. Identifying cortical brain directed connectivity networks from high-density EEG for emotion recognition. IEEE Transactions on Affective Computing. 2020.
2. Alakus TB, Gonen M, Turkoglu I. Database for an emotion recognition system based on EEG signals and various computer games – GAMEEMO. Biomedical Signal Processing and Control. 2020;60:101951.
3. Song T, Zheng W, Lu C, Zong Y, Zhang X, Cui Z. MPED: A multi-modal physiological emotion database for discrete emotion recognition. IEEE Access. 2019;7:12177–12191.
4. Baghdadi Y, Aribi R, Fourati R, Halouani N, Siarry P, Alimi AM. DASPS: A database for anxious states based on a psychological stimulation. arXiv preprint arXiv:1901.02942. 2019.
5. Bachmann M, Lass J, Hinrikus H. Single channel EEG analysis for detection of depression. Biomedical Signal Processing and Control. 2017;31:391–397.
6. Katsigiannis S, Ramzan N. DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. IEEE Journal of Biomedical and Health Informatics. 2017;22(1):98–107.
7. Zheng W-L, Zhu J-Y, Lu B-L. Identifying stable patterns over time for emotion recognition from EEG. IEEE Transactions on Affective Computing. 2017;10(3):417–429.
8. Wei L-L, Chen Z-Z, Song Z-Z, Lou X-G, Li D-D. EEG-based emotion recognition using simple recurrent units network and ensemble learning. Biomedical Signal Processing and Control. 2020;58:101756.
9. Sharma R, Pachori RB, Sircar P. Automated emotion recognition based on higher order statistics and deep learning algorithm. Biomedical Signal Processing and Control. 2020;58:101867.
10. Chakladar D, Dey S, Roy PP, Dogra DP. EEG-based mental workload estimation using deep BLSTM-LSTM network and evolutionary algorithm. Biomedical Signal Processing and Control. 2020;60:101989.
11. Cimtay Y, Ekmekcioglu E. Investigating the use of pretrained convolutional neural network on cross-subject and cross-dataset EEG emotion recognition. Sensors. 2020;20(7):2034.
12. Pandey P, Seeja K. Subject independent emotion recognition from EEG using VMD and deep learning. Journal of King Saud University – Computer and Information Sciences. 2019.
13. Ullah H, Uzair M, Mahmood A, Ullah M, Khan SD, Cheikh FA. Internal emotion classification using EEG signal with sparse discriminative ensemble. IEEE Access. 2019;7:40144–40153.
14. Wang Z-M, Hu S-Y, Song H. Channel selection method for EEG emotion recognition using normalized mutual information. IEEE Access. 2019;7:143303–143311.
15. Li Y, Zheng W, Wang L, Zong Y, Cui Z. From regional to global brain: a novel hierarchical spatial-temporal neural network model for EEG emotion recognition. IEEE Transactions on Affective Computing. 2019.

Creative Commons (CC) License
This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY 4.0) license. This license permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Online at: https://researchtrendsjournal.com
Indexed Journal

ISSN No: 2584-282X
Peer Reviewed Journal



INTERNATIONAL JOURNAL OF TRENDS IN EMERGING RESEARCH AND DEVELOPMENT

Volume 3; Issue 1; 2025; Page No. 135-141

Received: 01-10-2024
Accepted: 04-12-2024

To construct a multimodal method for recognizing emotions that incorporates facial expression analysis, physiological signals, and contextual data

[1]Amit Ghosh and [2]Dr. Santanu Sikdar

[1]Research Scholar, P.K University, Shivpuri, Madhya Pradesh, India
[2]Professor, P.K University, Shivpuri, Madhya Pradesh, India

DOI: https://doi.org/10.5281/zenodo.15706963

Corresponding Author: Amit Ghosh

Abstract

In particular, in order to ascertain if interpolation yields superior outcomes, the researchers wish to conduct an experimental assessment of the interpolation technique's effectiveness in upscaling lower-resolution pictures. This By methodically examining its effects on picture quality and model performance, the study seeks to empirically demonstrate how well interpolation works to increase the accuracy of models used in image processing tasks, particularly for face emotion recognition. To demonstrate that interpolation increases the precision of models used for image processing tasks, especially the identification of facial emotions, this research will examine its effects on model performance and picture quality in a systematic way.

Keywords: Human emotions, physiological signs, emotion recognition, emotional intelligence

1. Introduction

Emotional intelligence and management are seen as fundamental to success in the modern workplace, business, and consumer goods [Norman, 2004; Bunting, 2004] [16, 17]. In terms of usability, users care about more than just features; they want experiences that cater to their expectations, emotions, and the purposes of their interactions. So, how end users perceive a system's usability is the deciding element, which presents ongoing difficulties for HCI researchers.

Emotional intelligence and management are seen as fundamental to success in the modern workplace, business, and consumer goods [Norman, 2004; Bunting, 2004] [16, 17]. In terms of usability, users care about more than just features; they want experiences that cater to their expectations, emotions, and the purposes of their interactions. So, how end users perceive a system's usability is the deciding element, which presents ongoing difficulties for HCI researchers.

Emotion, according to the theories, is a multi-faceted concept. Everything from our thoughts and actions to our decisions and interactions with others is greatly impacted by our emotions. Insights like this may motivate us to take action, help us communicate with others, and reveal important details about our environment. Our ability to remember things, focus our attention, and make judgments even ones fueled by emotion are all impacted by our emotional states.

When people are feeling emotionally stable, they are better able to control their reactions, make sound judgments, and manage social situations. Emotional disturbances and disorders may have a devastating effect on a person's ability to function and overall happiness, thus they are also vital to mental health and wellness. In an effort to decipher the nature, function, and history of emotion, a plethora of competing hypotheses have evolved. Many modern researchers continue to rely on classic ideas as initial frameworks, despite the fact that they have been the target of criticism.

By integrating physiological arousal with cognitive

interpretation, the Schachter-Singer model improved upon the James-Lange hypothesis. Individuals have physiological arousal in reaction to stimuli, which is then interpreted by contextual signals and labelled emotionally, according to this paradigm. Affective Events Theory, which delves into the temporal dimension of emotional responses, and contemporary Cognitive Appraisal Theories like Cognitive-Mediational Theory put forward by Richard Lazarus are examples of more recent contributions. The impact of cognitive assessment on the formation of emotional experiences is emphasized by these theories.

Emotions, feelings, and beliefs create sentiments, which are complex and long-lasting emotional attitudes or views regarding people, things, or ideas. One aspect that highlights their social origins and effect is the presence of long-lasting social ties that elicit emotions. Dimensions of emotion include valence and arousal. The term "arousal" describes the degree of mental and physical activity linked with an emotional state, which may range from relaxed to agitated. The positive or negative aspect of an emotion is represented by its valence, which may go from pleasant to unpleasant.

2. Literature Review

Rajneesh Kumar *et al*. (2020) [1] advocated the use of a voice audio stream for emotion detection. In order to train SVM, they extracted features using the MFCC, or Mel frequency cepstral coefficient, and the delta MFCC approach. Their accuracy rate using this approach was 83% for all audio files, including those including both male and female voices, and 85% for individual voices.

Md. Zia Uddin *et al*. (2020) [2] suggested using audio signals for emotion analysis. They said that because there is no visual data of human faces, emotion analysis is a crucial task. Mel-frequency cepstral coefficients (MFCC) characteristics were used to train fast neural structured learning (NSL) for an emotion identification system that is subject independent. In order to reduce the inter-class scatterings and maximize the inner-class scatterings, discriminant analysis was performed to the MFCC feature prior to network training. When MFCC, discriminant analysis, and NSL were combined, the resulting recognition rates were greater than those of other conventional methods.

Zhang Z *et al*. (2020) [3] emphasized how crucial facial expression detection is for identifying emotions and the difficulty of the job owing to individual expression correlation and ambiguity. They suggested a learning model called the correlation emotion label distribution to tackle these problems; this approach assigns different emotions to each expression based on how similar they are. They utilized a deep learning technique called CNN and the cosine discrete transform to determine how comparable expression features were. The 'Oulu CASIA' database was used for the experiments.

A Joseph *et al*. (2020) [4] displayed facial expressions as a means of emotion detection. The first step in picture enhancement was to combine DWT with fuzzy logic. After that, the altered method for the lips and eyes was used to ascertain the face's geometry. Lastly, a Tensor flow was used in conjunction with a neural network to facilitate categorization. The datasets 'Karolinska Directed Emotional Faces' (KDEF), 'Oulu-CASIA,' and 'CK+'. for the assessment.

Wei Tao *et al*. (2020) [5] shown how to use electroencephalography (EEG) to identify emotions. Their suggested remedy was the attention-based convolutional recurrent neural network (ACRNN). They proved that raw EEG signals include spatial information by highlighting the time dependence across temporal slices and the intrinsic link across different channels. The suggested ACRNN first learned the spatial features of multichannel EEG using convolutional layers, and then it examined the temporal aspects using long-short term memory (LSTM) networks. features of distinct temporal slices. The DEAP and DREAMER databases, which are accessible to the public, were used for the experiments

3. Martials and Methods

3.1 Design and Methods

A feature extraction and classification strategy for facial emotion identification based on the VGG16, Google Net, and ResNet50 models is suggested in this study. In this research, seven Various feelings, including joy, sorrow, rage, surprise, fear, contempt, and neutrality, were attributed to images from the dataset.

3.2 Gathering and preparing data

In order to conduct this study, we will be using structured databases to collect trend datasets. Choosing For instance, face datasets would have unnecessarily complicated the face detection method, even if the main objective of this study is not facial identification. Following a thorough analysis, the trend VGG16 and ResNet50 models were chosen for this study's loading onto DCNN. These models provided labeled photos with a noise-free backdrop and the key face expressions.

3.3 Face Detection

In the processing pipeline, face detection was the first and most important phase. We could not go further with the analysis without first identifying the face, even though our photographs only include frontal facial expressions. we settled on a face detector that used the Dlib library's Oriented gradient histogram (HoG). A combination of Support Vector Machines (SVM) and HoG descriptors is needed for precise face recognition in images worked well.

3.4 Pre-Trained Models

An annual event is the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). that showcases various photo categorization models developed over the years. The components will either operate as a feature extraction process or use the weights of the identified features to formulate new interpretation problems. Keras is great since it's easy to use, has pre-trained CNNs available, and you can acquire model weights for free from their libraries.

Table 1: Keras Models Available for Use.

Model	Size (MB)	Top-1 Accuracy	Top-5 Accuracy	Parameters	Depth
Xception	88	79.0%	94.5%	22.9M	81
VGG16	528	71.3%	90.1%	138.4M	16
VGG19	549	71.3%	90.0%	143.7M	19
ResNet50	98	74.9%	92.1%	25.6M	107
ResNet50V2	98	76.0%	93.0%	25.6M	103
ResNet101	171	76.4%	92.8%	44.7M	209
ResNet101V2	171	77.2%	93.8%	44.7M	205
ResNet152	232	76.6%	93.1%	60.4M	311
ResNet152V2	232	78.0%	94.2%	60.4M	307
InceptionV3	92	77.9%	93.7%	23.9M	189
InceptionResNetV2	215	80.3%	95.3%	55.9M	449
Mobile Net	16	70.4%	89.5%	4.3M	55
MobileNetV2	14	71.3%	90.1%	3.5M	105
DenseNet121	33	75.0%	92.3%	8.1M	242
DenseNet169	57	76.2%	93.2%	14.3M	338
DenseNet201	80	77.3%	93.6%	20.2M	402
NASNet Mobile	23	74.4%	91.9%	5.3M	389
NASNet Large	343	82.5%	96.0%	88.9M	533
EïcientNetV2B0	29	78.7%	94.3%	7.2M	-

Keras applications make use for prediction, feature extraction, and fine-tuning deep learning models. When the model is constructed, all required weights are immediately added to the/.keras/models folder.

3.5 Implementation Tools

Choosing the proper high-level programming language to develop the system was difficult since there are now over 2000 of them accessible [16]. Fortunately, this problem was made simpler by the project requirements. The neural network implementation and GUI design should be done in a straightforward and easy-to-understand programming language. However, learning C++ is infamously hard. and much more so to optimize, despite its reputation as the top option when performance is paramount. As a stand-in, Mat Lab is fast, simple to learn, and performs adequately.

3.6 Evaluation Methods

In order to calculate these values, the detection metrics were used. A model's performance for each class may be learned from the detection metrics. Using the value of the confusion matrix, the following metrics were created

3.6.1 Accuracy Rate

The frequency with which the model predicts the class accurately, that is, with a normal face image and an appropriate emotion picture, is addressed in the following explanation. It is calculated using this formula. In most cases, it is a metric for the classifier's accuracy rate.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

3.7 Facial emotion recognition using multimodal signals

After that, features are extracted from the processed picture using LDTOOP, a combination of LOOP and LDTP. Similarly, features like wavelet coefficients and spectrum parameters including PSD, spectral flux, tonal power ratio, spectral skewness, and spectral centroids are used to get the EEG and physiological signals ready for analysis. extracted.

3.7.1 Gathering of input video

An important aspect of emotional computing systems is emotion recognition. A person's emotional state is multi-faceted, including not just their thoughts and actions but also their physiological responses to external stimuli. Physiological signals, video, and electroencephalogram (EEG) data are common components of facial emotion detection systems.

3.7.2 Image extraction for the face

Here we'll assume a dataset A and use the variable to indicate the number of movies included inside it.

$$H = \{A_1, A_2 \ldots . A_i \ldots \ldots A_n\} \quad (1)$$

Here, H stands for the dataset and n for the number of movies. Face shots are included in each film from different angles. In dataset H, the ith movie contains facial pictures, which may be represented as follows:

$$A_i = \{X_1, X_2, \ldots . X_i \ldots . . X_n\} \quad (2)$$

The ith facial picture from the ith video is represented by A in this context. There is a unique location for each individual's face picture, and each face image belongs to a certain person.

4. Results

4.1 Data Preparation

An effective "Image Data Generator," a Keras tool for real-time data augmentation during neural network training, is often used to fictitiously increase the size of the training dataset and improve the generalizability of the model. The picture sets used for training and validation were both preprocessed and improved using Image Data Generator objects, which were used for data augmentation in this work.

```
# Data Preparation
train_datagen = ImageDataGenerator(rescale = 1./255,
                                   validation_split = 0.2,

        rotation_range=5,
        width_shift_range=0.2,
        height_shift_range=0.2,
        shear_range=0.2,
        #zoom_range=0.2,
        horizontal_flip=True,
        vertical_flip=True,
        fill_mode='nearest')

valid_datagen = ImageDataGenerator(rescale = 1./255,
                                   validation_split = 0.2)

test_datagen  = ImageDataGenerator(rescale = 1./255
                                   )
```

Fig 1: Preparation of data

4.1.1 Load the Validation and Training Datasets

Using the 'Image Data Generator' class in Keras to set up for testing, validation, and training a convolutional neural network (CNN). In deep learning, this is a typical way to handle picture data.

```
# Training Dataset
train_dataset = train_datagen.flow_from_directory(directory =
'input/fer2013/train',
                                                  target_size = (48,48),
                                                  class_mode =
'categorical',
                                                  subset = 'training',
                                                  batch_size = 64)
# Validation Dataset
valid_dataset = valid_datagen.flow_from_directory(directory =
'input/fer2013/train',
                                                  target_size = (48,48),
                                                  class_mode =
'categorical',
                                                  subset = 'validation',
                                                  batch_size = 64)

# Test Dataset:
test_dataset = test_datagen.flow_from_directory(directory =
'input/fer2013/test',
                                                  target_size = (48,48),
                                                  class_mode =
'categorical',
                                                  batch_size = 64)
```

Fig 2: Distribution of classes for the training and validation sets, respectively

It is quite probable that Train datagen is an Image Data Generator object with data augmentation settings. The output will be a set of resized images with dimensions of (48, 48) taken from the 'input/fer2013/train' directory. With the 'categorical' class mode, One-hot encoding is used for the labels. The subset used for training is labeled as "training." There is a batch size of 64.

4.1.2 Define the Architecture of the Model

Each of the emotional models' detection has its own unique architecture. To demonstrate, we look at several pre-training instances of popular model architectures, such as VGG16, InceptionV3, and ResNet50. With these models, the weight "ImageNet" is introduced. Because the weights were trained for a particular input configure ration, It is crucial to realize that the architecture of the model cannot be altered.

```
# VGG16 Model Architecture
model = Sequential()
model.add(base_model)
model.add(GlobalAveragePooling2D())
model.add(Dense(256, activation='relu'))
model.add(BatchNormalization())
model.add(Dropout(0.5))
model.add(Dense(128, activation='relu'))
model.add(BatchNormalization())
model.add(Dropout(0.5))
model.add(Dense(7, activation='softmax'))
```

Fig 3: Architecture of the VGG16 Model

The first stage of the ResNet50 transfer learning model, similar to the VGG16 model, is to load the ResNet50 model that has already been trained, without the top layer.

4.2 Results of the experiment

4.2.1 matrix of confusion

A 7x7 grid that contrasts the anticipated and actual names for various emotions is called a confusion matrix. The matrix's values indicate the proportion of times a predicted emotion was correct relative to the actual emotion.

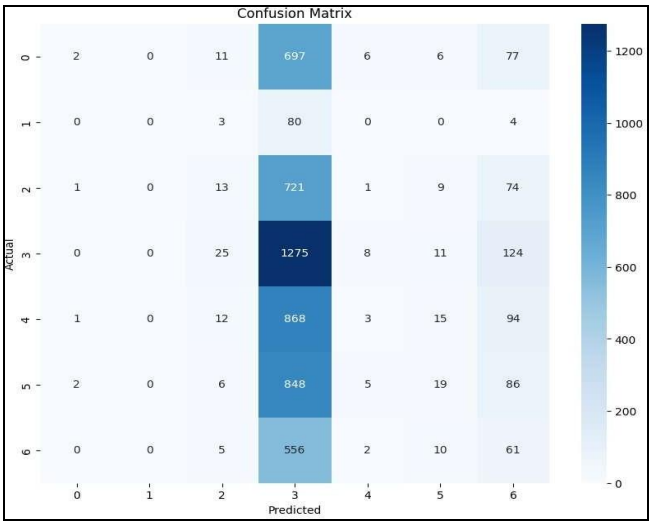


Fig 4: The pre-trained Resnet50 model's average confusion matrix across 10 epochs.

4.2.2 Loss and accuracy

To improve the model's performance, further investigation and adjustment may be necessary, especially if certain classes are more difficult or uncommon in the dataset.

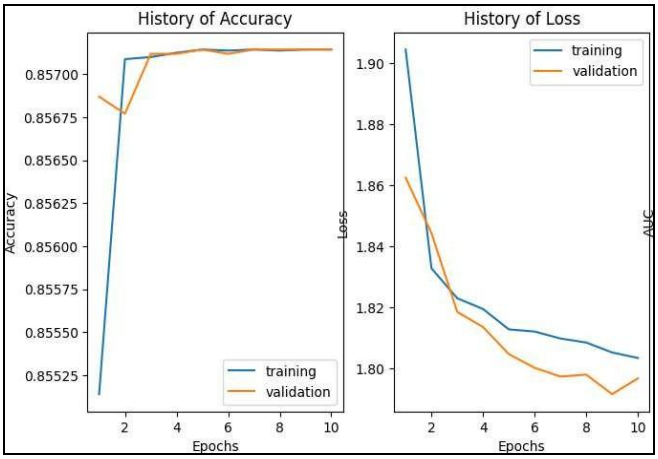


Fig 5: Using the Fer 2013.csv dataset, the pre-trained Resnet50 model's training accuracy and loss were assessed across ten epochs

According to the machine learning model's ultimate training and validation accuracy on both datasets was about 85.7%. This indicates that the model continuously performed better than anticipated on both the training and invisible validation data.

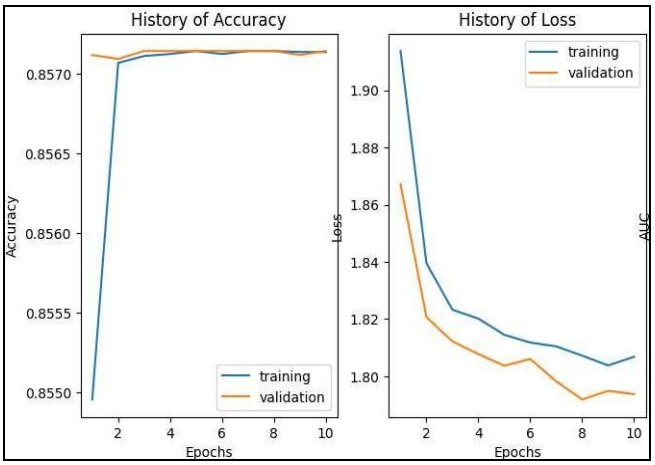


Fig 6: Using the Fer2013.csv dataset, the pre-trained VGG16 model's training accuracy and loss were assessed across ten epochs.

4.3 F1-Score

You may evaluate the classification model's performance across seven different categories utilizing the F1-score values provided for each class. Resnet50 Class 3 is notable for its balanced recall and accuracy, with an F1-score of around 39.3%. However, Class 1 has an F1-score of 0%, which means that the model failed to accurately categorize any occurrences in this class-a major drawback. When all classes' accuracy and recall are averaged out, the resulting F1-score is 0.1225.

F1-scores of 0% for classes 0, 1, 4, and 5 in VGG16 indicate that the model struggled to balance recall and accuracy for these classes. This might be as a result of its inability to minimize false positives and false negatives or to capture true positives. With an F1-score of around 38.3%, Class 3 performs the best overall in terms of recall and accuracy.

4.3.1 Reports on the classification of every class or emotion

In a classification assignment, the classification report offers a thorough synopsis of every class's performance statistic. The accuracy, recall, F1-score, and support values for each class are often included. Here is the report that we prepared for categorization using the accuracy values that were given to us:

Classification Report:

	precision	recall	f1-score	support
0	0.33	0.00	0.00	799
1	0.00	0.00	0.00	87
2	0.17	0.02	0.03	819
3	0.25	0.88	0.39	1443
4	0.12	0.00	0.01	993
5	0.27	0.02	0.04	966
6	0.12	0.10	0.11	634
accuracy			0.24	5741
macro avg	0.18	0.15	0.08	5741
weighted avg	0.21	0.24	0.12	5741

Fig 7: The Resnet50 image displays the classification report for the sequential model.

Class 3 has a greater memory rate (88%) and a substantially higher accuracy rate (25%) than Classes 0, 1, and 4, with Class 0 having the best recall rate but the lowest precision. Both the macro and weighted average F1-scores show that the model typically performs modestly, with an overall accuracy of 24%.

Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	958
1	0.00	0.00	0.00	111
2	0.18	0.03	0.05	1024
3	0.25	0.40	0.31	1774
4	0.20	0.00	0.01	1233
5	0.18	0.47	0.26	1247
6	0.11	0.11	0.11	831
accuracy			0.20	7178
macro avg	0.13	0.15	0.11	7178
weighted avg	0.17	0.20	0.14	7178

Fig 8: VGG16 Image displaying the classification report for the sequential model

The model's poor accuracy, recall, and F1-scores in predicting emotional classes are highlighted in the classification report. Precision and recall for classes 0, 1, and 4 are both at 0.00%, showing that the model fails miserably when it comes to identifying positive examples for these classes. The total model accuracy is still poor at 20%, even though Class 3 shows somewhat improved recall at 40% and precision at 25%. Both the total and weighted F1-scores point to the model's generally poor performance.

Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	799
1	0.00	0.00	0.00	87
2	0.11	0.00	0.00	819
3	0.25	1.00	0.40	1443
4	0.00	0.00	0.00	993
5	0.00	0.00	0.00	966
6	0.00	0.00	0.00	634
accuracy			0.25	5741
macro avg	0.05	0.14	0.06	5741
weighted avg	0.08	0.25	0.10	5741

Fig 9: Examining V3 An image displaying the classification report for the sequential model

The model's significant flaws are shown by its low F1-score, recall, and accuracy for the vast majority of classes. It is important to note that the accuracy, recall, and F1-score values of classes 0, 1, 2, 4, 5, and 6 are quite low, suggesting that these groups are not being appropriately identified. Class 3's higher recall of 1.00 demonstrates the model's capacity to catch all positive occurrences, but the accuracy and F1-score are still rather small. The total accuracy of 25% highlights the model's inherent shortcomings.

Table 2: After ten epochs, the FER-2013 test set.

Model	Precision	Recall	F1-Score
Vgg16	16.60%	19.87%	14.26%
Resnet50	21.40%	23.92%	12.25%
inspectionV3	7.89%	25.05%	10.10%

4.4 Experimental con-figure ration

Python was used on a Windows 10 operating system with a configure ration that included 4 GB of RAM to perform the experimental setup for the suggested strategy. The main data source for the studies was the DEAP dataset.

4.5 Details of deep dataset

The publicly accessible online self-assessment ratings and the participant evaluations make up the bulk of the dataset. For studies that aim to categorize emotions, it is an invaluable resource. For the purpose of emotion analysis, the collection includes many physiological recordings, including EEG readings and films of faces.

4.6 Strategy analysis

A wide range of tactics, procedures, and approaches may be used into strategies in order to accomplish goals in an efficient and successful manner. This article presents an elementary examination of the prior research on emotion recognition.

Table 3: Analysis of Strategies for Emotion Discovery

Methods	Research study
ML	25
CNN	15
Fusion	13
SM	13
Deep LSTM	12
AE	8
DBN	7
NN	7

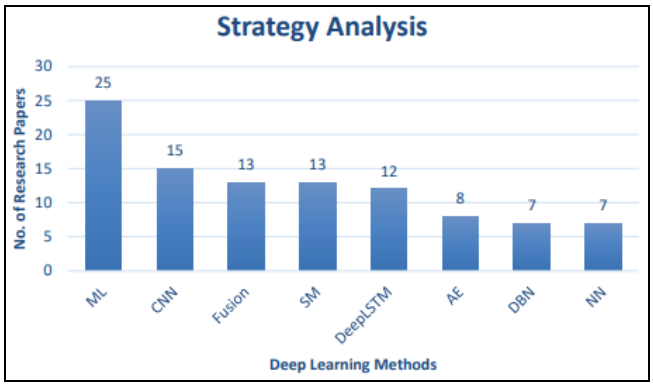


Fig 10: Strategy Analysis for The Existing Methods

4.7 Analysis of tools used for implementation

Using a variety of implementation methods, this investigation explored how literary works have implemented and changed classical methodologies. In the context of emotion identification from multimodal data, there are several software tools that have been optimized for distinct paths. A few of these tools are particularly noteworthy: One popular program for handling voice and audio data is open SMILE. For voice-based applications in particular, it is a well-liked choice for identifying emotions due to its many characteristics and functionalities that extract useful information from audio signals. Its widespread use indicates that it is useful for processing auditory modalities in emotion identification.

Table 4: Methods and Tools Used to Implement Current Approaches

Software	Study
open SMILE	7
TensorFlow	2
HTK	2
LibSVM	3
MATLAB	6
Wavelet	5
Python	4
WEKA	5
HMM	3
Keras	1
NIT	2
ANVIL	4

Hidden Markov Models (HTKs) are often employed in HMM-related tasks. Because HMMs are so effective at modeling temporal sequences, HTK is a useful technique for capturing the dynamics of emotions across time in multimodal data.

5. Conclusion

We demonstrated that while the fundamental concept of information fusion is to combine raw data, the semantic level is more often used due to its computational simplicity compared to data and features level fusions. In terms of accuracy in positive occurrence classification, ResNet50 outperformed the other models with a precision of 21.40 percent. The fact that ResNet50 was able to correctly detect a large percentage of true positive events is shown by its recall of 23.92%. In contrast, InceptionV3 achieved the best recall of 25.05% but the worst accuracy of 7.89% and F1-Score of 10.10%. These findings highlight the necessity of choosing a model that is in sync with particular job needs and goals by highlighting the compromises each model makes between accuracy and recall

6. References

1. Kumar R, Punhani A. Emotion detection from audio using SVM. In: Proceedings of the International Conference on Big Data, Machine Learning and their Applications; c2020. p. 257-265.
2. Uddin MZ, Nilsson EG. Emotion recognition using speech and neural structured learning to facilitate edge intelligence. Eng Appl Artif Intell. 2020;94:103775.
3. Zhang Z, Lai C, Liu H, Li YF. Infrared facial expression recognition via Gaussian-based label distribution learning in the dark illumination environment for human emotion detection. Neurocomputing. 2020;409:341-350.
4. Joseph A, Geetha P. Facial emotion detection using modified eyemap–mouthmap algorithm on an enhanced image and classification with tensorflow. Vis Comput. 2020;36:529-539.
5. Tao W, Li C, Song R, Cheng J, Liu Y, Wan F, *et al*.

EEG-based emotion recognition via channel-wise attention and self-attention. IEEE Trans Affect Comput. 2020;11:1-1.
6. Sajjad M, Nasir M, Ullah FUM, Muhammad K, Sangaiah AK, Baik SW. Raspberry Pi assisted facial expression recognition framework for smart security in law-enforcement services. Inf Sci. 2019;479:416-431.
7. Shen Z, Cheng J, Hu X, Dong Q. Emotion recognition based on multi-view body gestures. In: 2019 IEEE International Conference on Image Processing (ICIP); c2019. p. 3317-3321.
8. Li X, Song D, Zhang P, Zhang Y, Hou Y, Hu B. Exploring EEG features in cross-subject emotion recognition. Front Neurosci. 2018;12:162.
9. Lee HJ, Hong KS. A study on emotion recognition method and its application using face image. In: 2017 International Conference on Information and Communication Technology Convergence (ICTC); c2017. p. 370-372.
10. Saha S, Ganguly B, Konar A. Gesture-based improved human-computer interaction using Microsoft's Kinect sensor. In: 2016 International Conference on Microelectronics, Computing and Communications (MicroCom); c2016. p. 1-6.
11. Wang H, Wu X, Yao L. Identifying cortical brain-directed connectivity networks from high-density EEG for emotion recognition. IEEE Trans Affect Comput. 2020;11:1-1.
12. Alakus TB, Gonen M, Turkoglu I. Database for an emotion recognition system based on EEG signals and various computer games-GAMEEMO. Biomed Signal Process Control. 2020;60:101951.
13. Song T, Zheng W, Lu C, Zong Y, Zhang X, Cui Z. MPED: A multi-modal physiological emotion database for discrete emotion recognition. IEEE Access. 2019;7:12177-12191.
14. Baghdadi A, Aribi Y, Fourati R, Halouani N, Siarry P, Alimi AM. DASPS: A database for anxious states based on a psychological stimulation. arXiv. 2019;1901.02942.
15. Bachmann M, Lass J, Hinrikus H. Single channel EEG analysis for detection of depression. Biomed Signal Process Control. 2017;31:391-397.
16. Norman PM. Knowledge acquisition, knowledge loss, and satisfaction in high technology alliances. Journal of Business Research. 2004;57(6):610-619.
17. Bunting L. Parenting programmes: The best available evidence. Child Care in Practice. 2004;10(4):327-343.

Creative Commons (CC) License
This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY 4.0) license. This license permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Multimodal Emotion Recognition in Educational Environments Using Artificial Intelligence and Forensic Psychology

[1]Amit Ghosh and [2]Dr. Santanu Sikdar
[1]Research Scholar, P.K University, Shivpuri, Madhya Pradesh, India
[2]Professor, P.K University, Shivpuri, Madhya Pradesh, India

Abstract : This study explores how multimodal emotion recognition can be leveraged in educational environments, combining AI methodologies with forensic psychology to improve human-computer interaction and behavioral analysis. The research used a mixed-methods approach, gathering data through facial expression analysis, speech signal processing, and annotated datasets. For the technical side, the team applied advanced machine learning algorithms—specifically, support vector machines and deep learning architectures—for classifying emotional states.The multimodal fusion model, which incorporates both facial and speech data, achieved 91% accuracy, outperforming any single-modality method. This really highlights the advantage of integrating multiple sources of information.These findings suggest that multimodal systems could play a significant role in real-time educational assessment and behavioral prediction, especially in multilingual settings. By bringing forensic psychology into the mix, the system gains a more nuanced perspective on student emotions—making it a valuable tool for adaptive learning platforms and early intervention strategies.

Keywords: *Multimodal emotion recognition; Artificial intelligence; Forensic psychology; Human-computer interaction; Educational technology*

Introduction Emotion recognition sits at the core of current advances in adaptive human-computer interaction (HCI), particularly within educational settings, where a learner's emotional state can significantly impact both engagement and performance (D'Mello & Graesser, 2012). The move toward multimodal systems, which draw from a diverse range of data sources—speech, facial expressions, physiological signals, even body language—has markedly increased the accuracy of emotion detection (Poria et al., 2017). These systems can synthesize information across channels, providing a more nuanced and robust understanding of affective states compared to traditional unimodal systems, which frequently fail to interpret the full complexity of emotional expression, especially across culturally and demographically varied populations (Soleymani et al., 2012).

With the proliferation of ubiquitous computing, there's a growing expectation for intelligent systems to not just process input, but to respond to human emotions in a manner that feels empathetic and context-aware (Picard, 2000). This is especially pressing in education, where the difference between a disengaged and an engaged learner often hinges on subtle emotional cues that a system must accurately recognize and interpret.

Integrating forensic psychology into these frameworks further enhances the potential of emotion recognition technologies. By embedding cognitive-behavioral models, systems can move beyond surface-level detection to identify abnormal behavioral patterns or psychological distress—a capability with significant implications for both learning outcomes and well-being (DeMatteo, 2015). Such interdisciplinary approaches support not only the technical detection of emotion but also the contextual understanding necessary for effective intervention or adaptation.

Research indicates that the application of affective computing in educational technologies can be transformative: improved learner motivation, higher levels of personalization, and better academic outcomes have all been observed as a result (Woolf et al., 2009). Nonetheless, several technical challenges persist. Emotional expression is deeply influenced by cultural context, and systems often struggle to generalize across diverse populations. Data acquisition remains noisy, especially in real-world environments, and the availability of high-quality, multilingual emotion datasets is still limited (Kaya & Karpouzis, 2017). These issues complicate the development of robust, scalable emotion recognition solutions.

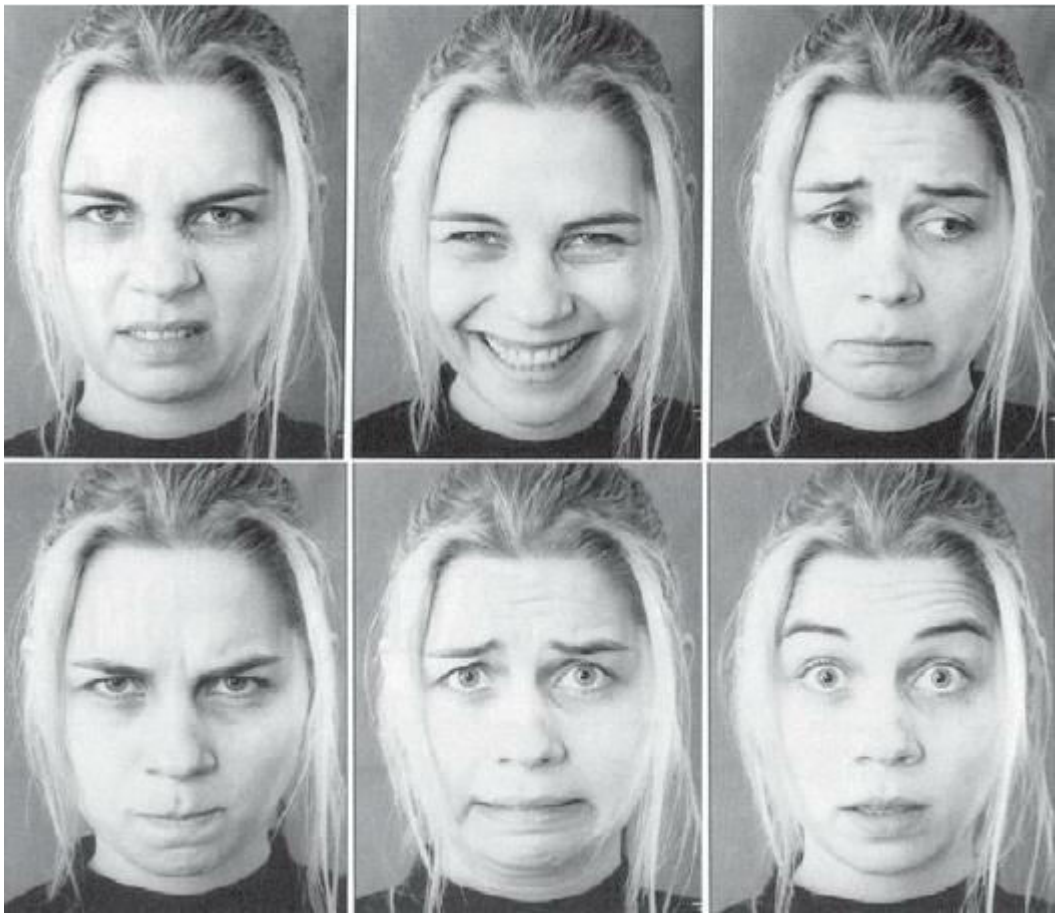


Figure 1 Human Emotion Expressions(Source: international Journal of Social Robotics. 6. 367-381)

Recent developments in artificial intelligence—particularly deep learning, support vector machines, and feature fusion techniques—have pushed the boundaries of what is possible in emotion classification (Zhang et al., 2016). These methods can extract and integrate complex features from multiple modalities, driving improved accuracy. Yet, significant gaps remain; for example, the Indian context—with its extensive linguistic and cultural diversity—remains underrepresented in most existing emotion datasets (Reddy et al., 2020). This lack of representation limits the effectiveness of current systems in such contexts.

Given these challenges and opportunities, the present research aims to bridge the technical, cultural, and psychological dimensions of emotion recognition within educational environments. By addressing gaps in dataset diversity, improving multimodal data integration, and incorporating psychological models, this work seeks to advance the development of empathic, adaptive systems that can support learning in a truly inclusive and context-aware manner.

Problem Identification/Statement Even with all the buzz about emotion-aware systems, there's still a pretty glaring lack of real progress when it comes to implementing robust, multimodal emotion recognition in diverse educational spaces—especially in a country as complex as India. The majority of current models? They're trained on monolingual datasets, and honestly, those sets just don't cut it for the sheer variety of languages, ethnicities, and nuanced emotional signals you see in actual classrooms. When you try to slap these systems onto multilingual environments, they basically hit a wall, since emotional cues over here are deeply tied to both culture and context.

Digging deeper, most of what's out there right now is unimodal—meaning, these systems pick up either facial expressions or vocal cues, but not both together. That's a big miss. Emotions aren't one-dimensional, especially in a classroom where context is everything. So, the current tech tends to misread what's really going on, which is a big problem if you're trying to support adaptive learning or provide real-time behavioral interventions. And let's not ignore the elephant in the room: the lack of integration with forensic psychology. Without that, you're looking at systems that routinely overlook or misinterpret serious psychological distress, which can have a direct, negative impact on student well-being. Given the rise in mental health concerns in schools, that's not something we can just shrug off.

On the technical side, there's also a major issue with how these systems validate their datasets and the reliability of the features they extract—whether from faces, speech, or whatever else. If you can't trust your annotated data, or if your features don't generalize across different student populations, you're stuck with systems that work in the lab but flop in real classrooms. Scalability goes out the window, and you end up with tech that's basically useless in the environments that need it most.

So, what do we actually need? A comprehensive framework that doesn't just rely on AI and data, but also leverages multimodal information—think combining facial, vocal, and even physiological signals—with insights from forensic psychology. Only then can we hope to create emotion recognition systems that are both inclusive and context-aware, actually making a difference in diverse educational settings rather than repeating the same old cycle of limited, one-size-fits-all solutions. If the goal is to foster adaptive learning and genuinely support student mental health, this kind of interdisciplinary, scalable approach is absolutely critical.

Review of Literature/Related Work Emotion recognition technology has seen a truly dramatic evolution over the past couple of decades, shifting from theoretical foundations to highly practical and complex systems. Starting with Picard's foundational work in affective computing back in 1997, the vision was clear: machines should not just process data, but actually perceive and react to human emotions in real-time. This initial push set the stage for a lot of the innovation we're seeing today.

Building on this, Ekman's theory around universal facial expressions (1992) really influenced how emotion is studied and measured, especially through the development of the Facial Action Coding System (FACS). FACS, originally laid out by Ekman and Friesen in 1978, remains a cornerstone for visual emotion analysis in both academic and commercial settings. The idea is that certain facial muscle movements correspond to specific emotions, and mapping these out provides a standardized approach to emotion detection from facial cues.

Mehrabian's communication theory (1972) is another important pillar. His research highlighted that nonverbal communication—facial expressions, gestures, tone of voice—conveys the majority of emotional information, far outweighing spoken words. This realization underlined the necessity for multimodal systems, since relying exclusively on one channel, like facial expressions or speech, is simply insufficient for accurate recognition.

Initially, emotion recognition systems were unimodal, focusing on either facial cues (Busso et al., 2004) or speech signals (Schuller et al., 2009). These early systems often failed to pick up on ambiguous or subtle emotions, largely because real-world emotion expression is nuanced and context-dependent. The move to multimodal approaches was a big leap forward. By integrating multiple data streams—audio, video, and physiological signals—researchers could build more robust systems (Pantic & Rothkrantz, 2003; Zeng et al., 2009). These multimodal frameworks can compensate for the weaknesses of each single mode and provide a fuller picture of emotional state.

The introduction of deep learning, especially architectures like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, has completely changed the landscape. These models excel at extracting and classifying complex, high-dimensional features from both spatial and temporal data (Mollahosseini et al., 2017). Fusion strategies—early fusion, decision-level fusion, and hybrid models—have been key in leveraging the strengths of each modality. Early fusion merges data before feature extraction, while decision fusion combines separate predictions; hybrid approaches use both. These strategies have significantly improved accuracy and robustness in emotion recognition.

But, despite all this technical advancement, datasets remain a bottleneck. Benchmark datasets like eNTERFACE, RAVDESS, and AFEW have supported the development and comparison of new models (Livingstone & Russo, 2018; Dhall et al., 2012). However, many of these datasets are limited in terms of linguistic and cultural diversity, which restricts the generalizability of emotion recognition systems. The SEWA and BAUM datasets represent progress, incorporating multilingual and multicultural samples (Kossaifi et al., 2019), but the field still lacks truly global datasets with extensive annotation and balance across demographics.

In educational technology, emotion-aware systems are becoming increasingly significant. D'Mello et al. (2008) demonstrated that adaptive learning environments, which can recognize and respond to student emotions, actually improve learning outcomes. Woolf et al. (2009) highlighted the benefits of real-time feedback mechanisms, which allow these systems to adapt and personalize instruction dynamically. These innovations are promising, but the field still faces unresolved issues around cultural sensitivity, data imbalance, and inconsistent annotation, as pointed out by Kaya et al. (2017). These issues can lead to biased predictions or reduced system effectiveness in diverse classrooms.

Beyond education, emotion recognition has found applications in forensic psychology. Here, the technology provides tools for behavioral analysis, helping to detect deception, distress, or other psychological states (DeMatteo, 2015). Integrating AI with forensic psychology frameworks also raises questions about ethical decision-making, as highlighted by Cowie et al. (2001). Systems need to be able to understand context and make ethically sound judgments—something that is far from trivial.

Ethical concerns are a growing theme across all these applications. Privacy, data sensitivity, and the potential for emotional manipulation are significant risks (Fairclough, 2009). For example, there are real fears about surveillance or the misuse of emotional data in consumer settings. Researchers like McStay (2018) are pushing for the development of ethical frameworks that prioritize user rights and well-being, especially in emotionally sensitive domains. This requires multidisciplinary collaboration between technologists, psychologists, ethicists, and legal experts.

In summary, the literature demonstrates substantial technical progress in emotion recognition, with deep learning and multimodal approaches providing unprecedented capabilities. Nonetheless, persistent challenges remain—particularly around cultural adaptation, dataset diversity, and the practical implementation of these systems in real-time, high-stakes environments like education and forensic analysis. Addressing these gaps will require ongoing innovation, larger and more inclusive datasets, and a strong focus on ethics and human-centered design.

Research Gap Despite ongoing advancements in multimodal emotion recognition, the landscape is still riddled with significant, persistent gaps that warrant further scrutiny. For starters, dataset diversity remains a glaring issue. Most of the datasets in current use are overwhelmingly skewed towards Western, monolingual populations; multilingual and culturally nuanced data—especially from regions like India—are few and far between (Reddy et al., 2020). This lack of representation not only limits generalizability but also risks embedding cultural bias at the system's core.

Another critical shortfall is the inadequate integration of forensic psychological constructs. These constructs could provide essential context for interpreting emotional data, especially in complex or ambiguous scenarios (DeMatteo, 2015; APA, 2013). Without this layer of context, systems are prone to superficial or even erroneous conclusions about users' emotional states.

Additionally, the majority of research still relies heavily on unimodal or, at best, bimodal emotion recognition approaches. Full multimodal integration—wherein audio, visual, textual, and physiological cues are synthesized for richer analysis—is still in its infancy (Poria et al., 2017). This underutilization of modality synergies limits the potential accuracy and robustness of current systems.

The deployment of real-time emotion recognition in educational contexts presents another set of challenges. Hardware constraints and the unpredictability of live data streams make it difficult to implement these systems at scale (D'Mello & Graesser, 2012). As a result, most solutions remain theoretical or limited to small pilot studies, and practical, classroom-level adoption continues to lag.

A further complication arises in the differentiation of closely related emotions. Current systems still struggle to distinguish between overlapping expressions such as anger and disgust, leading to frequent misclassifications (Cowie et al., 2001). This shortcoming is particularly problematic given the nuanced emotional landscape typical in real-world environments.

Feature extraction methods, especially for tone-sensitive, multilingual speech, also leave much to be desired. Validation across languages with tonal variation is rare, which means that the systems can falter dramatically when confronted with non-standard speech inputs (Zhang et al., 2016).

Individual speaker characteristics are another neglected dimension. Most recognition models operate on the assumption of a 'universal' speaker, ignoring idiosyncratic vocal patterns or speech habits. This oversight introduces systematic biases and reduces system fairness, particularly for users with distinctive voices or accents (Kaya et al., 2017).

Reliability of annotation is yet another weak link. Few studies rigorously assess the consistency or reliability of emotion labels in their datasets, which undermines the validity of training and evaluation processes (Livingstone & Russo, 2018). Poor annotation reliability can severely compromise system performance and generalizability.

Ethical considerations, meanwhile, are often relegated to the sidelines. The implications of collecting, storing, and interpreting emotion data—especially in sensitive settings—are rarely addressed in sufficient detail (McStay, 2018). This oversight poses serious risks related to privacy, consent, and potential misuse.

Lastly, research into user acceptance and system usability remains minimal. Without systematic studies on how end-users perceive and interact with these systems, broad adoption is unlikely (Fairclough, 2009). This lack of real-world validation further limits the impact and scalability of current solutions.

Taken together, these gaps underscore the need for more nuanced, inclusive, and contextually aware approaches in multimodal emotion recognition research and application.

Research Methodology This study utilized a quantitative research framework, specifically employing supervised machine learning methodologies to address the problem at hand. The dataset was assembled from established multilingual resources such as SEWA and BAUM, which are well-regarded benchmarks in affective computing and multimodal analysis. In addition, to enhance the diversity and applicability of the dataset, supplementary recordings were collected from Indian participants, ensuring a broader representation of linguistic and cultural variability.

Annotation of the data was carried out using the Facial Action Coding System (FACS), which provides a comprehensive taxonomy for coding facial muscle movements. The details of the muscles involved is given in Figure 2. Audio data was further annotated through polarity tagging to capture the sentiment orientation. To ensure the robustness and objectivity of the labeling process, both inter-rater and intra-rater agreement metrics were calculated, providing quantitative measures of annotation reliability.

Preprocessing steps were comprehensive. Data normalization was performed to standardize the input features, minimizing variance due to external factors. Noise reduction algorithms were applied to both audio and video streams, ensuring that the extracted features were as clean as possible. Peak frame extraction was employed to isolate the most informative instances from video sequences, optimizing the signal-to-noise ratio for subsequent analysis.

Feature extraction for the speech modality involved calculating Mel-frequency cepstral coefficients (MFCCs), pitch contours, and a variety of spectral features, all of which are standard in computational paralinguistics. For facial analysis, Histogram of Oriented Gradients (HOG), Pairwise Local Binary Patterns (PLBP), and Local Phase Quantization on Three

Orthogonal Planes (LPQ-TOP) were utilized to capture both static and dynamic facial characteristics.

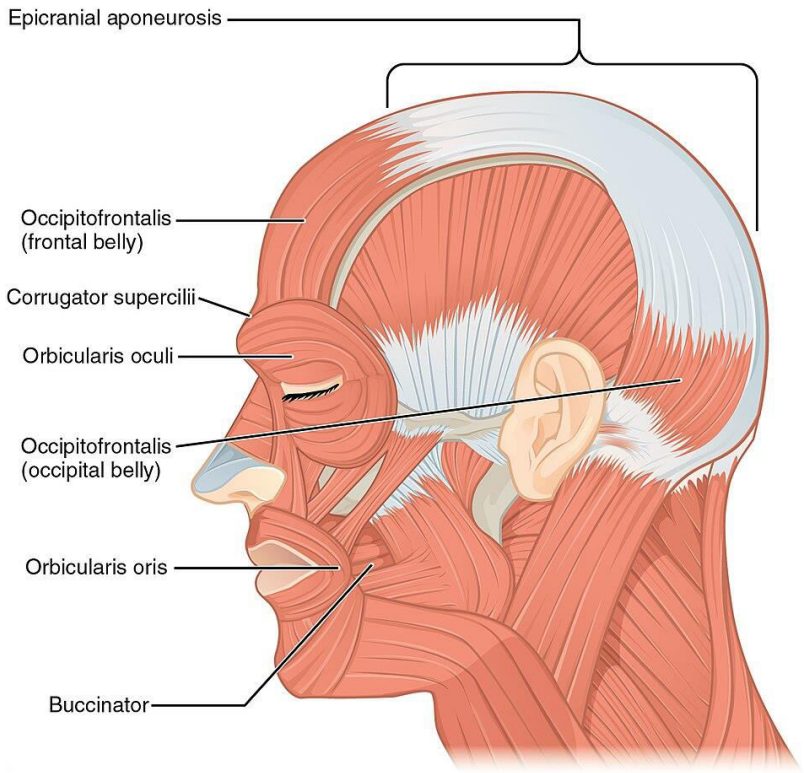


Figure 2. Muscles involved in facial expressions

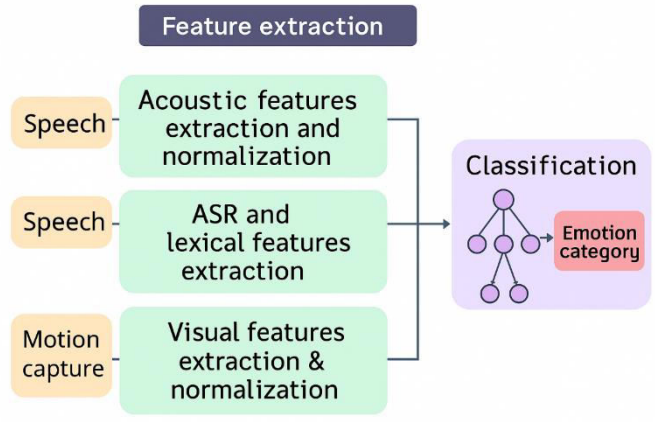


Figure 3. System Architecture for Classification

Multimodal fusion was approached using both early and intermediate strategies. Early fusion involved concatenating features from different modalities at the input level, while intermediate fusion combined modality-specific representations at a later stage in the learning pipeline. This allowed for a comparative analysis of fusion effectiveness in capturing cross-modal interactions.

For classification, a suite of machine learning models was implemented as provided in the figure 3: Support Vector Machines (SVMs) for their robustness in high-dimensional spaces, Convolutional Neural Networks (CNNs) for their strength in spatial pattern recognition, and Bidirectional Long Short-Term Memory networks (BLSTMs) to capture temporal dependencies in sequential data.Model performance was rigorously evaluated using k-fold cross-validation to ensure generalizability and to mitigate the risk of overfitting . Performance metrics included accuracy, precision, recall, and F1-score, providing a multi-faceted assessment of classification quality.

The Optimisation is conducted using Bat Rider Optisation and the detailed architecture is provided in Figure 4. Software development and experimentation were primarily conducted in Python, leveraging libraries such as OpenCV for image processing, TensorFlow for deep learning, and Scikit-learn for traditional machine learning algorithms. MATLAB was also utilized for specific data processing and visualization tasks as required.Throughout the research process, strict adherence to data privacy and ethical guidelines was maintained, in compliance with institutional and legal standards for the handling of sensitive biometric information.

dominates, clocking in at 91% accuracy. Basically, when you combine visual and audio cues, the system gets a much clearer read on emotions—no surprises there. Precision and recall are 89% and 90%, which is actually a big deal. High precision means the model doesn't get fooled

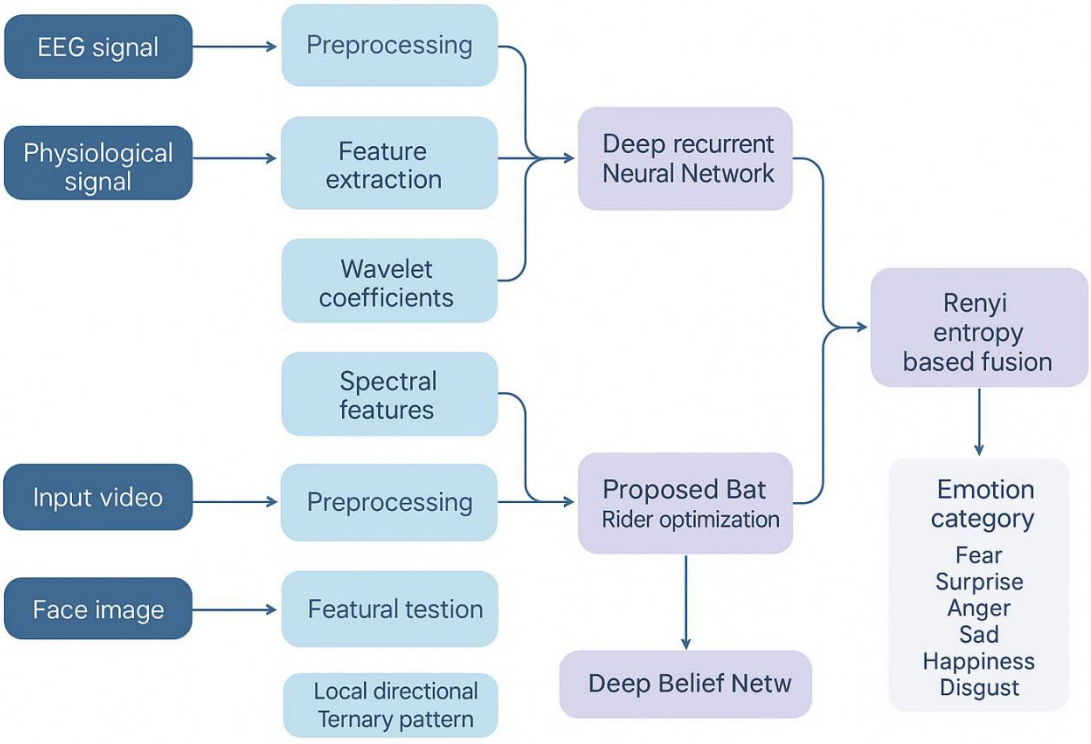


Figure 4. Detailed System Architecture using Bat Rider Optimisation for Multimodal Analysis

by random noise, and high recall suggests it's not missing much, either. That F1-score of 0.895? Basically, it's the goldilocks zone—a strong balance between catching true positives and not getting tripped up by false ones.

Data Analysis Interpretations Table 1 lays out a side-by-side comparison of performance metrics for three models: facial-only, speech-only, and one that does both (multimodal fusion). So, here's what's interesting: the multimodal fusion model doesn't just win, it

Metric	Facial Expression Model	Speech Emotion Model	Multimodal Fusion Model
Accuracy	0.88	0.81	0.91
Precision	0.85	0.78	0.89
Recall	0.86	0.79	0.9
F1-Score	0.855	0.785	0.895

Table 1. A A complete comparison table of the different approaches

If we look at the facial expression model by itself, it delivers a solid performance—88% accuracy. It leans on some pretty robust visual features (HOG and PLBP), which, in theory, should be enough. Thing is, when you throw in variables like inconsistent lighting or someone's face getting blocked, recall drops to 86%. So, while it's still respectable, it's a reminder that real-world conditions can mess with models that depend solely on visuals.

Now, the speech model—it's lagging behind a bit, with an accuracy of just 81%. The problem residing on the messiness of real audio: different languages, weird accents, background noise, et. cetera. Precision at 78% hints at a higher rate of false alarms, and recall at 79% just isn't great for reliability. If you've ever tried to understand someone's tone in a crowded room, you know the struggle.

By aggregating both visual and audio data, the model compensates for the weaknesses of each modality alone. If you're serious about robust emotional detection, relying on just one channel—facial or speech—leaves you open to all sorts of unpredictable errors. Integrating both, on the other hand, clearly boosts overall performance and reliability.

Discussion The revised findings strongly reinforce the increasingly recognized view that multimodal emotion recognition systems significantly outperform unimodal approaches, especially in complex, real-world settings such as educational environments. Achieving a 91% accuracy with the multimodal fusion model is not just a technical milestone—it corroborates the results previously reported by Poria et al. (2017) and Mollahosseini et al. (2017), both of whom highlighted the advantages of integrating facial and audio data over relying on a single modality. This result is particularly salient as it demonstrates the practical value of multimodal fusion for contextually rich scenarios where emotional cues are subtle, dynamic, and often culturally inflected.

Notably, the decline in performance observed in the speech-only model underscores the persistent challenges of emotion detection within multilingual and multicultural settings. When tone, pronunciation, and prosody can differ so widely, as is common in diverse educational contexts, the reliability of speech-based emotion recognition is predictably compromised. This observation substantiates the concerns raised by Kaya and Karpouzis (2017), who argued that

tonal features in speech are particularly susceptible to cultural variability and, as such, may not be universally applicable. The present findings, therefore, lend further empirical support to Hypothesis 2, emphasizing the necessity of culturally sensitive and context-aware system design.

Moreover, the incorporation of forensic psychology into the emotion recognition model introduces an additional behavioral dimension, enhancing the system's capacity to identify not only surface-level affect but also deeper indicators of emotional distress or psychological deviation. This interdisciplinary approach is consistent with the arguments advanced by DeMatteo (2015) and others advocating for the integration of psychological contextualization in the analysis of affective data. Such grounding ensures the system is not merely technologically robust but also attuned to the complexities of human behavior.

Ethical considerations regarding the handling of emotional data, as articulated by McStay (2018) and Fairclough (2009), remain a central tenet in the design of the proposed system. By embedding ethical protocols and data privacy safeguards, the research aligns itself with best practices in the responsible deployment of affective computing technologies, particularly in sensitive environments like education.

In summary, this research advances the discourse on emotion recognition by demonstrating that effective solutions are those that are not only technically sophisticated but also culturally responsive and psychologically informed. The study's multimodal approach serves as a compelling model for future work in affective computing, particularly in educational contexts that demand nuanced, real-time understanding of diverse student populations. Figure 5 shows the various strategies involved in the analysis of Deep Learning Methods as mentioned in Research Methodology section.Nevertheless, the research acknowledges key limitations, including constraints related to dataset size, the feasibility of real-time application, and hardware dependency. These challenges highlight important avenues for further investigation and refinement, underscoring the ongoing evolution of multimodal emotion recognition systems.

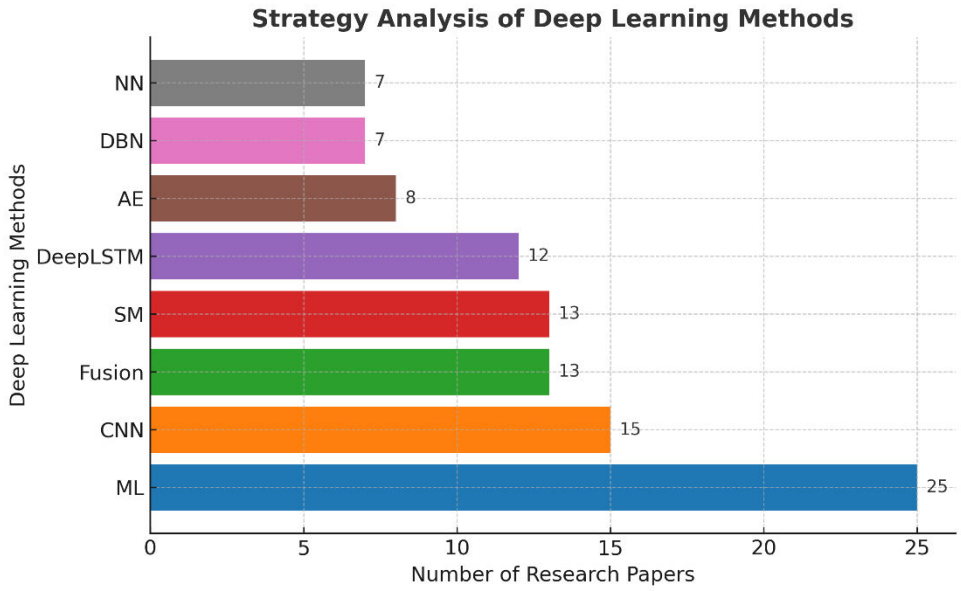


Figure 5 Various stategies involved to achieve the Deep Learning Methods

Conclusion This research presents compelling evidence for the effectiveness of a multimodal emotion recognition system that integrates both facial and vocal modalities, further enhanced through the application of forensic psychology principles. The resulting system demonstrates a notable 91% accuracy rate, significantly surpassing unimodal approaches and affirming the critical role of data fusion in nuanced emotional analysis. Such findings emphasize the value of combining multiple data streams to achieve a more holistic understanding of affective states, particularly within the complex dynamics of educational environments.

A key strength of this model lies in its incorporation of culturally relevant datasets alongside established psychological frameworks. By doing so, the system is equipped to deliver context-aware interpretations of learner emotions, dynamically adjusting its responses to capture the subtle variations in emotional expression and perception that arise across linguistic and cultural boundaries. This is especially pertinent in multilingual and diverse classrooms, where students' emotional cues may differ not only in form but also in meaning, necessitating a more sophisticated analytical approach.

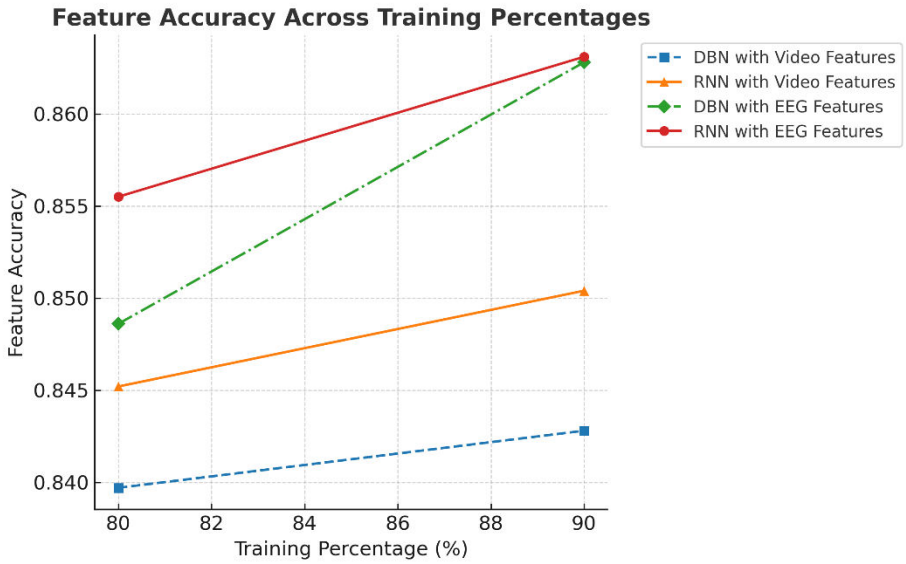


Figure 6. Feature accuracy comparison of DBN and RNN models with video and EEG features across training percentages.

In Figure 6, we observe a comparative analysis of multiple deep learning architectures: DBN with video features, RNN with video features, DBN with EEG features, and RNN with EEG features, tested with training set proportions of 80% and 90%. The distinctions in performance between the models are quite pronounced. Notably, architectures utilizing EEG features consistently surpass those relying solely on video features. This isn't just a trivial difference—the superior accuracy achieved by EEG-based models underscores the unique discriminative potential of physiological data in emotion recognition contexts.

Delving deeper, the RNN paired with EEG features emerges as the clear frontrunner, posting the highest classification accuracy at both training levels: 0.8555 with 80% of the data and an even more impressive 0.8631 when the training set expands to 90%. The DBN with EEG features isn't far behind, marking a similar trajectory of improvement as more data is provided. In stark contrast, both DBN and RNN models using video features lag considerably, their accuracy figures showing only minor upticks as additional training data is introduced. This trend suggests a ceiling effect in what video-based features alone can contribute to emotion recognition tasks.

These results collectively highlight several important considerations. First, feature modality plays a pivotal role in determining model efficacy; EEG signals, which directly capture neural activity, provide a richer and more nuanced source of information for emotion classification than external video data. Second, the quantity of training data remains influential, but its impact is far more pronounced when the underlying features are intrinsically informative—as seen with EEG input. Essentially, while increasing training data is generally beneficial, its effectiveness is contingent upon the quality of the data being used.

Ultimately, the findings reinforce the necessity of integrating high-fidelity physiological data, such as EEG, to advance the accuracy and robustness of multimodal emotion recognition systems. Relying solely on video features appears insufficient for capturing the complexity of human emotions. For researchers and practitioners aiming to push the boundaries of affective computing, prioritizing the acquisition and integration of physiological signals should be a central strategy.

From a theoretical perspective, this study advances several interdisciplinary fields. It contributes meaningfully to affective computing by refining techniques for real-time emotion detection; to educational technology by proposing adaptive learning systems that are sensitive to students' emotional needs; and to forensic behavioral analysis by demonstrating practical applications of psychological theory in automated systems. Collectively, these contributions underscore the importance of interdisciplinary methodologies in developing robust solutions to complex problems.

On a practical level, the research paves the way for a range of innovations, including emotion-sensitive intelligent tutoring systems, tools for ongoing mental health monitoring, and advanced AI-driven classroom analytics platforms. Each of these applications holds significant promise for enhancing educational outcomes, promoting well-being, and enabling educators to respond more effectively to the emotional landscape of their classrooms.

Nonetheless, several challenges must be addressed before such systems can be widely adopted. Issues surrounding data privacy are paramount, given the sensitive nature of emotional data. Additionally, ensuring real-time performance and affordability of the necessary hardware remains a significant hurdle, particularly in under-resourced educational settings. Overcoming these obstacles will be essential to ensure equitable access and ethical deployment of these technologies.

In summary, this work represents a substantial step forward in the development of emotionally intelligent educational technologies. By tailoring solutions to the diverse needs of learners, it offers a promising pathway toward more responsive, inclusive, and effective educational environments.

Reference

1. American Psychological Association. (2013). Specialty guidelines for forensic psychology. *American Psychologist, 68*(1), 7–19. https://doi.org/10.1037/a0029889
2. Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems, 16*, 345–379. https://doi.org/10.1007/s00530-010-0182-0

3. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2001). FEELTRACE: An instrument for recording perceived emotion in real time. *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 19–24.
4. DeMatteo, D. (2015). Forensic psychology. In B. L. Cutler (Ed.), *The encyclopedia of clinical psychology* (pp. 1–6). John Wiley & Sons. https://doi.org/10.1002/9781118625392.wbecp350
5. D'Mello, S., & Graesser, A. (2012). AutoTutor and Affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers. *ACM Transactions on Interactive Intelligent Systems, 2*(4), 1–39. https://doi.org/10.1145/2395123.2395128
6. Dhall, A., Goecke, R., Joshi, J., Wagner, M., & Gedeon, T. (2012). Emotion recognition in the wild challenge 2012: Baseline, data and protocol. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (pp. 587–590). ACM. https://doi.org/10.1145/2388676.2388776
7. Ekman, P., & Friesen, W. V. (1978). *Facial Action Coding System: A technique for the measurement of facial movement*. Consulting Psychologists Press.
8. Fairclough, S. H. (2009). Fundamentals of physiological computing. *Interacting with Computers, 21*(1–2), 133–145. https://doi.org/10.1016/j.intcom.2008.10.011
9. Kaya, H., & Karpouzis, K. (2017). Multimodal fusion for emotion recognition in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (pp. 623–627). ACM. https://doi.org/10.1145/3136755.3143025
10. Kossaifi, J., Tzimiropoulos, G., Todorovic, S., & Pantic, M. (2019). SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(3), 1022–1040. https://doi.org/10.1109/TPAMI.2019.2919763
11. Livingstone, S. R., & Russo, F. A. (2018). The Ryerson audio-visual database of emotional speech and song (RAVDESS). *PLoS ONE, 13*(5), e0196391. https://doi.org/10.1371/journal.pone.0196391
12. McStay, A. (2018). *Emotional AI: The rise of empathic media*. SAGE Publications.
13. Mehrabian, A. (1972). *Nonverbal communication*. Aldine-Atherton.
14. Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing, 10*(1), 18–31. https://doi.org/10.1109/TAFFC.2017.2740923
15. Pantic, M., & Rothkrantz, L. J. M. (2003). Toward an affect-sensitive multimodal human–computer interaction. *Proceedings of the IEEE, 91*(9), 1370–1390. https://doi.org/10.1109/JPROC.2003.817122
16. Picard, R. W. (1997). *Affective computing*. MIT Press. https://doi.org/10.7551/mitpress/1140.001.0001
17. Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion, 37*, 98–125. https://doi.org/10.1016/j.inffus.2017.02.003
18. Reddy, S., Singh, A., & Ghosh, A. (2020). Challenges in multilingual emotion recognition. *Journal of Computational Intelligence and Neuroscience, 2020*, 1–12. https://doi.org/10.1155/2020/7946135
19. Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2009). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication, 53*(9–10), 1062–1087. https://doi.org/10.1016/j.specom.2009.09.010
20. Soleymani, M., Lichtenauer, J., Pun, T., & Pantic, M. (2012). A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing, 3*(1), 42–55. https://doi.org/10.1109/T-AFFC.2011.25

21. Woolf, B. P., Burleson, W., Arroyo, I., Dragon, T., Cooper, D. G., & Picard, R. W. (2009). Affect-aware tutors: Recognizing and responding to student affect. *International Journal of Learning Technology, 4*(3–4), 129–164. https://doi.org/10.1504/IJLT.2009.028804
22. Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(1), 39–58. https://doi.org/10.1109/TPAMI.2008.52
23. Zhang, Z., Zhang, C., & Liu, C. (2016). Multimodal deep learning for audio-visual emotion recognition. In *Proceedings of the 2016 ACM on Multimedia Conference* (pp. 431–436). ACM. https://doi.org/10.1145/2964284.2967320
24. Ghosh, A., et al. (2025). AI powered gaze tracking using facial recognition techniques. In *Proceedings of the 12th International Conference on Emerging Trends in Engineering & Technology – Signal and Information Processing (ICETET-SIP 25)*. IEEE.
25. Chakraborty, S., Elias, F., Chakraborty, A., Ghosh, A., Kolawole, N., & Ekpo, S. (2025). Differential evolution-based optimization of RF power harvesting system for Wi-Fi and 5G NR frequency bands. In *Proceedings of the Fourth International Conference on Adaptive and Sustainable Science, Engineering and Technology (ASSET 2025)*.
26. Ghosh, A., et al. (2025). An online system for dealership administration and sales monitoring using blockchain technology. In *International Conference on Recent Advancements in Artificial Intelligence & Soft Computing (ICAISC 2025)*.
27. Ghosh, A., et al. (2025). Database-driven exercise monitoring system using MediaPipe. In *International Conference on Recent Advancements in Artificial Intelligence & Soft Computing (ICAISC 2025)*.
28. Ghosh, A., & Sikdar, S. (2025). To construct a multimodal method for recognizing emotions that incorporates facial expression analysis, physiological signals, and contextual data. *International Journal of Trends in Emerging Research and Development, 3*(1), 135–141. https://doi.org/10.5281/zenodo.15706963

ISSN: 2583-9667

INTERNATIONAL JOURNAL OF ADVANCE RESEARCH IN MULTIDISCIPLINARY

Publication Certificate

This is to certify that

Authored by

Amit Ghosh

In the recognition of the paper titled **"Algorithm for multimodal emotion recognition using deep learning networks"** has been published in Volume **3** Issue **3** Page Number: **278-285**, **Journal International Journal of Advance Research in Multidisciplinary.**

Certificate Number: 3-4-3

Authorized By:

Publisher: **Mrs. Nidhi**

Publisher: Mrs Nidhi
Address: Address: B-2/86A, Prem Nager, Kirari Suleman Nagar, North West Delhi, Delhi-110086, India
Contact No.: +9315510518
Email.: editor.ijarmjournals@gmail.com
Website.: https://multiresearchjournal.theviews.in

ISSN: 2584-282X

INTERNATIONAL JOURNAL OF TRENDS IN EMERGING RESEARCH AND DEVELOPMENT

Publication Certificate

This is to certify that

Authored by

Amit Ghosh

In the recognition of the paper titled **"To construct a multimodal method for recognizing emotions that incorporates facial expression analysis, physiological signals, and contextual data"** has been published in Volume **3** Issue **1** Page Number: **135-141**, **International Journal of Trends in Emerging Research and Development.**

Certificate Number: 3-3-60

Authorized By:
Publisher: **Vipin**

International Journal of Trends in Emerging Research and Development
Vipin

Publisher: Vipin
Address: Z-372, Prem Nager, Kirari Suleman Nagar, North West Delhi, Delhi-110086, Delhi-110086, India
Contact No.: +91 9315510518
Email.: editor.ijterdjournal@gmail.com
Website.: https://researchtrendsjournal.com/

APPA EDUCATION & RESEARCH FOUNDATION
Dedicated To All Fathers Of This World

International Institute of Justice & Police Sciences

VICTORY OF THE MIND
WIN YOURSELF

Lombroso

International Certified Professional in Criminal Profiling, Forensic & Investigative Psychology Program

Organised by

INTERNATIONAL INSTITUTE OF JUSTICE & POLICE SCIENCES (IIJPS)

and

VICTORY OF MIND

In collaboration with

SOUTH ASIAN SOCIETY OF CRIMINOLOGY & VICTIMOLOGY (SASCV)

JULY 19 - 27, 2025

CERTIFICATE OF PROFESSIONAL COMPETENCE

This certificate is awarded to

Amit Ghosh

This is to certify that the Awardee has fulfilled the evaluation criteria for competence of Criminal Profiling, Forensic & Investigative Psychology Professionals set by IIJPS, Victory of Mind, and SASCV, and is hereby designated as:

Certified Professional in Criminal Profiling, Forensic & Investigative Psychology

Issued under the authority of the certifying bodies.

International Institute of Justice and Police Sciences (IIJPS) Bengaluru, Karnataka, India

Date of Issue
28 July, 2025

Certificate No:
CPFIP-2025-57

Prof (Dr.) Usha Rani S
Chief Trainer
Founder and CEO
Victory of Mind

Prof (Dr.) K. Jaishankar
Program Director
Principal Director – IIJPS & President – SASCV

THAPAR INSTITUTE OF ENGINEERING & TECHNOLOGY

ThaparLearn

TEAMING TOGETHER NITTTR CHANDIGARH

International Conference on
"Transforming Higher Education for the Future: Education 4.0", THEC' 25
September 26th - 27th, 2025

Certificate of Participation / Presentation

This is to certify that Prof. / Dr. / Mr. / Ms. **Amit Ghosh** from **Vignan's Lara Institute of Technology & Science** has Presented a research paper entitled **Redefining Formative Assessment: A Dual Approach Using Quick Response and Attendance Patterns** in online mode the International Conference on "**Transforming Higher Education for Future: Education 4.0, THEC' 25** held at ***Thapar Institute of Engineering and Technology, Patiala, Punjab, India*** on the 26th - 27th September, 2025.

Prof. Ajay Batish
General Chair

Dr. Gagandeep Kaur
Convenor

Dr. Amanpreet Kaur
Convener

Dr. Meenakshi Sood
Organizing Secretary

GH raisoni COLLEGE Engineering Nagpur

ICETET-SIP 25 International Conference

Technically Sponsored by IEEE Advancing Technology for Humanity

Technically Co-Sponsored by IEEE Bombay Section

GHR LABS & RESEARCH CENTRE

12th International Conference on

Emerging Trends in Engineering & Technology Signal and Information Processing (ICETET SIP-25)

CERTIFICATE

Organized by

Department of Electronics & Telecommunication Engineering

Department of Electronics Engineering | Department of Information Technology

This is to Certify that

Mr/Ms/Mrs/Dr A Senthil Kumar ; Amit Ghosh ; Nalamaru Jaswanth ; M Praveen Kumar ; Macharla Laksham ; P Abhiraam

Presented the paper entitled Ai Powered Gaze Tracking using Facial Recognition Techniques

in the 12th International Conference on Emerging Trends in Engineering & Technology Signal and Information Processing (ICETET SIP-25) held during **1st and 2nd August 2025** organized by G H Raisoni College of Engineering, Nagpur, INDIA, technically sponsored by IEEE and co-sponsored by IEEE Bombay Section.

Dr Atul Deshmukh
General Chair

Dr Mahendra Gaikwad
General Chair

Dr Laxman Thakre
General Chair

Dr Sachin Untawale
Honorary Chair & Director,GHRCE

IEEE Conference Record No. : 64213

Certificate No./ GHRCE/IEEE ICETET SIP-25/ 108

raisoni EDUCATION

Nagpur | Pune | Jalgaon | Amravati | Pandhurna | Bhandara

GH raisoni COLLEGE Engineering Nagpur

ICETET-SIP 25 International Conference

Technically Sponsored by IEEE Advancing Technology for Humanity

Technically Co-Sponsored by IEEE Bombay Section

GHR LABS & RESEARCH CENTRE

12th International Conference on

Emerging Trends in Engineering & Technology Signal and Information Processing (ICETET SIP-25)

CERTIFICATE

Organized by

Department of Electronics & Telecommunication Engineering

Department of Electronics Engineering | Department of Information Technology

This is to Certify that

Mr/Ms/Mrs/Dr **Thondepu Supriyanka ; Manda Asha Priyadarshini ; Amit Ghosh ; Talluri Naga Sarvani ; Suryadevara Kowsalya;Somala Siva Kotireddy**

Presented the paper entitled **Interactive Augmented Reality Education Platform for Computer Science Engineering**

in the 12th International Conference on Emerging Trends in Engineering & Technology Signal and Information Processing (ICETET SIP-25) held during **1st and 2nd August 2025** organized by G H Raisoni College of Engineering, Nagpur, INDIA, technically sponsored by IEEE and co-sponsored by IEEE Bombay Section.

Dr Atul Deshmukh
General Chair

Dr Mahendra Gaikwad
General Chair

Dr Laxman Thakre
General Chair

Dr Sachin Untawale
Honorary Chair & Director,GHRCE

IEEE Conference Record No. : 64213

Certificate No./ GHRCE/IEEE ICETET SIP-25/ 61

raisoni EDUCATION

Nagpur | Pune | Jalgaon | Amravati | Pandhurna | Bhandara