Facial Emotion Recognition Using Deep Learning to Identify the Problems Related to Mental Health

A dissertation submitted to the

Department of Electrical and Electronic Engineering

University of Dhaka

In partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

E.E.E.

Submitted by:

Fakir Mashuque Alamgir

Registration no. **85/2019-2020**

Academic Session: **2019-2020**

Under the supervision of

Professor Dr. Md. Shafiul Alam

Department of Electrical and Electronic Engineering

University of Dhaka, Dhaka-1000, Bangladesh

September 2024.

Certification

This is to certify that the thesis entitled “Facial Emotion Recognition Using Deep Learning to Identify the Problems Related to Mental Health” is submitted by Fakir Mashuque Alamgir in September 2024 in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the department of Electrical and Electronic Engineering, University of Dhaka. This work is based on his work under my supervision.

I, with this, declare that this submission is the work of Fakir Mashuque Alamgir. To the best of my knowledge, it contains no materials previously published or written by another person or substantial proportions of material that have been accepted for the award of any other degree or diploma at the University of Dhaka or any educational institution, except where due acknowledgment is made in the thesis. Any contribution made to the research by others with whom Mr. Fakir Mashuque Alamgir has worked at the University of Dhaka or elsewhere is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of Mr. Fakir Mashuque Alamgir's research work, except to the extent that assistance from others is acknowledged.

....

Supervisor

Professor Dr. Md. Shafiul Alam

Department of Electrical and Electronic Engineering

Declaration

I do, at this moment, declare that the submitted thesis entitled “Facial Emotion Recognition Using Deep Learning to Identify the Problems Related to Mental Health” has been composed by me, and all the works presented here are of my findings.

I further declare that this work has not been submitted anywhere for any degree/diploma or other academic award.

Fakir Mashuque Alamgir

Registration no. 85/2019-2020

Academic Session: 2019-2020

Abstract

Facial emotions, such as neutral, happiness, sadness, fear, annoyance, anger, and surprise, are interpreted similarly across different cultures. These expressions indicate a person's emotional state, signal approval or disapproval of others' behaviors in social situations, and reveal mental or neurological disorders, if any. With the advent of high computing systems, image processing, and deep learning algorithms in the recent decade, facial expression recognition (FER) technology has achieved high accuracy in recognizing emotions from facial images. In many applications, the current technologies have achieved better performance than humans. However, identifying emotions from facial images under constrained circumstances remains challenging due to obstructions, poor or improper lighting conditions, and different positions of the head/body. Moreover, identifying mental or neurological disorders requires knowledge of the problem domain, appropriate deep-learning models, and labeled data sets. This research aims to enhance the performance of facial expression/emotion recognition models and apply them to real-world scenarios, such as identifying autism spectrum disorder in children, which is increasing rapidly, particularly in the post-COVID era.

The major contributions of this thesis present several novel classifier methods for identifying and discriminating facial emotions of individuals into seven different emotions: neutral, happiness, sadness, fear, annoyance, anger, and surprise, as endorsed by the American Psychological Association in 2008. Individuals with strong emotional well-being may still have a range of physical and mental issues. The proposed deep learning-based Artificial intelligence models generate sophisticated decisions using real-time data, a key strength that is highly relevant in identifying mental health problems.

Emotion recognition is complicated in children with Autism spectrum disorder (ASD) when they are suffering from speech issues and social communication issues, which leads to the requirement of an effective emotion recognition approach. Motivated by the alarming facts of autism spectrum disorder globally and locally, this thesis presents a unique dual-branch CNN-based visual transformation model to identify special children with a binary classifier. Since there is not a single dataset available for special children in terms of our country and south Asia, a real-time dataset was created with the high-resolution images that were collected from the two special schools of

Dhaka, Bangladesh. An original dataset was created for future researchers, which is considered to be one of the key contributions of this research program. The outcome of the developed model is finally evaluated with the other approaches and models, outperforming existing techniques in accuracy and other measures and successful detection of special children.

The proposed multiple approaches have been evaluated, considering precision, recall, f-measure, accuracy, specificity, and recognition rate. The proposed models demonstrated superior performance in all measures compared to the other algorithms and successfully detected special children. We also believe the proposed model will benefit our country's underprivileged people who have limitations in the early detection of mental health problems. Thus, the objectives of this Ph.D. research will be worthwhile.

Table of Contents

List of Figures

List of Tables

This thesis work is dedicated.

To

My beloved parents, wife, and children.

Acknowledgment

It is my great pleasure that Allah, the Almighty, who is merciful and the most gracious, gave me the patience to travel a long journey of Ph.D. research and allowed me to complete my Doctoral Thesis.

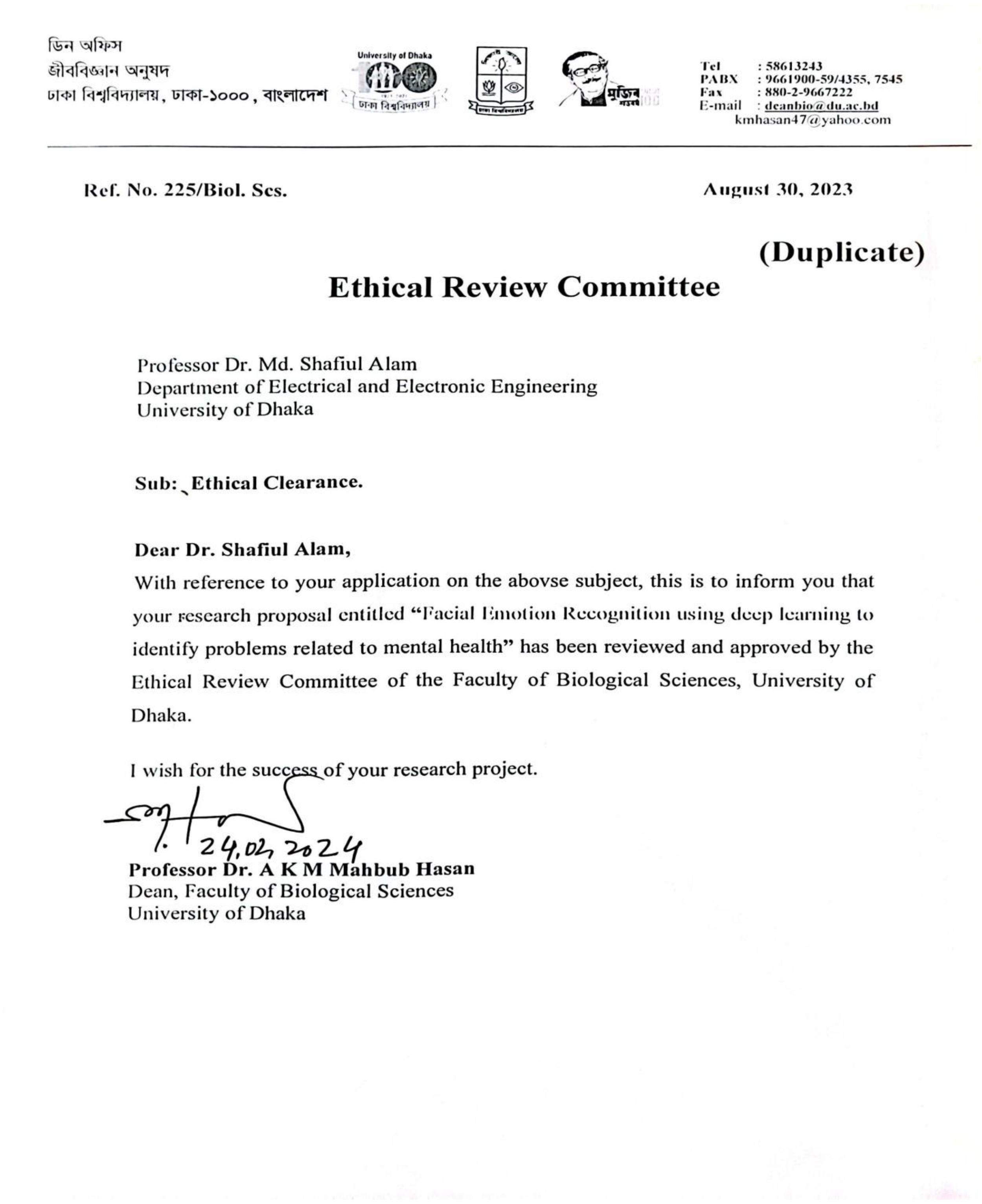
I express my deepest gratitude to my supervisor, Professor Dr. Md. Shafiul Alam, Professor, Department of Electrical and Electronic Engineering, University of Dhaka, for his extraordinary support and supervision. It is my rare privilege to get an opportunity to work with such a prudent, devoted, knowledgeable, inspiring, kind, and extraordinary researcher. I cannot express my appreciation for his continuous extraordinary support, invaluable advice, relentless encouragement, and unwavering guidance throughout my Ph.D. program.

Also, I want to thank Professors Abul Kalam Azad, Mosabber Uddin Ahmed, and Saeed Mahmud Ullah for their support during my coursework.

Finally, completing this research would not be possible without psychological motivation, continuous support, and creating a comfortable environment for my family members. In particular, my mother and my late father-in-law pushed me up and continuously encouraged me to complete my Ph.D. research. In addition, my wife, Tania Sultana, continuously busted me up and encouraged and helped me to create a calm research environment at home to complete my research. Love goes to my lovely children, Uzaan and Uzmaa, who have a meaningful influence on my life. In addition, I also want to mention my deceased father, who was the real backbone of my whole educational journey.

Fakir Mashuque Alamgir

Ethical Clearance Certificate



ডিন অফিস
জীববিজ্ঞান অনুষদ
ঢাকা বিশ্ববিদ্যালয়, ঢাকা-১০০০, বাংলাদেশ

Tel : 58613243
PABX : 9661900-59/4355, 7545
Fax : 880-2-9667222
E-mail : deanbio@du.ac.bd
kmhasan47@yahoo.com

Ref. No. 225/Biol. Scs. **August 30, 2023**

(Duplicate)

Ethical Review Committee

Professor Dr. Md. Shafiul Alam
Department of Electrical and Electronic Engineering
University of Dhaka

Sub: Ethical Clearance.

Dear Dr. Shafiul Alam,

With reference to your application on the abovse subject, this is to inform you that your research proposal entitled "Facial Emotion Recognition using deep learning to identify problems related to mental health" has been reviewed and approved by the Ethical Review Committee of the Faculty of Biological Sciences, University of Dhaka.

I wish for the success of your research project.

24,02,2024

Professor Dr. A K M Mahbub Hasan
Dean, Faculty of Biological Sciences
University of Dhaka

Abbreviation

AERS	Automatic expression recognition system
GSNMF	Graph-preserving sparse nonnegative matrix factorization
CNN	Convolutional Neural Networks
DBN	Deep Belief Networks
SVM	Support Vector Machine
FER	Facial Emotion Recognition
ACNN	Attention mechanism Convolutional Neural Networks
RAF	Real-world Affective Face Collection
EM	Expectation-Maximization
DLP-CNN	Deep locality-preserving convolutional neural network
LDTP	Local directional ternary pattern
HCL	Human-Computer Interaction
CS-LBP	Center-symmetric local binary pattern
GMM	Gaussian mixture model
CS	confidence score
DSAE	Deep sparse autoencoders
DRP	Digital Rock Physics
CT	Computed Tomography
SDAEN	Stacked denoising auto-encoders network
GAN	Generative adversarial networks
MRI	Magnetic resonance imaging
HDR	High Dynamic Range
PCA	Principal component analysis

SIFT	Scale-invariant feature transform
BEMD	Bidimensional empirical mode decomposition
FUFE	Flexible unsupervised feature extraction
GWO	Gray wolf optimizer
BGWO	Binary grey wolf optimization
PSO	Particle swarm optimization
OBIA	Object-Based Image Analysis
VHR	Very36 High-Resolution
CSSA	Chaotic slap swarm algorithm
COS	Classification Optimization Score
ACO	Ant Colony Optimization
BPSO	Bare bones particle swarm optimization
WOA	Whale Optimization Algorithm
ALO	Ant Lion Optimizer
IBGSA	Improved Binary Gravitational Search Algorithm
MICCA	Multi set integrated canonical correlation Analysis
MFCC	Mel frequency cepstral coefficient
FFT	Fast Fourier transform
DCT	Discrete Cosine Transform
JAFFE	Japanese Female Facial Expression
KDEF	Karolinska Directed Emotional Faces

Chapter 1: Introduction

1.1 Introduction

Imagine a world where a simple glance can reveal the depths of a person's struggles. Facial emotion recognition, powered by deep learning, is not just a technological marvel; it's a lifeline. Every furrowed brow, every twitch of a lip tells a story, a silent scream for help. Algorithms examine through countless images, learning to detect the subtle nuances of human expression, transforming pixels into profound insights about mental health.

In the evolving landscape of mental health care, the integration of technology has prompted innovative approaches to identify and address mental health issues. Among these advancements, Facial Emotion Recognition (FER) using deep learning stands out as a promising tool. This technology employs complex algorithms to analyze facial expressions, offering insights into emotional states that can be pivotal in the early detection and management of mental health disorders. While some critics raise valid concerns regarding privacy, accuracy, and the complexity of mental health, the potential benefits of FER technology cannot be overlooked. As we delve into the arguments and counterarguments surrounding this topic, it becomes evident that facial emotion recognition can revolutionize mental health care, facilitating early detection, improving treatment accuracy, and enhancing accessibility to services.

FER can provide early detection of mental health issues, which is crucial in addressing the growing mental health crisis globally. The cornerstone of effective mental health treatment is timely intervention. By identifying emotional distress through facial cues, FER technology can alert caregivers and professionals to potential problems before they escalate into more severe conditions. For instance, a study utilizing FER in clinical settings demonstrated that patients exhibiting signs of anxiety or depression could be flagged for further evaluation and immediate support. This proactive approach not only aids in reducing the severity of mental health disorders but also fosters a culture of awareness and prevention.

Moreover, FER systems can track emotional changes over time, creating a dynamic profile of an individual's mental health. This longitudinal data tracking enables healthcare providers to recognize patterns and anomalies in emotional expression, thereby enhancing their capacity to tailor interventions to individual needs. In essence, the use of FER technology can serve as an early warning system, ensuring that individuals receive the necessary support before their mental health deteriorates further.

Deep learning models significantly enhance the accuracy of emotion recognition, a critical factor in the efficacy of mental health interventions. Unlike traditional methods, advanced algorithms in deep learning are designed to analyze subtle facial cues that may go unnoticed by the human eye. For example, slight variations in the positioning of facial muscles can indicate underlying emotional states, such as anxiety or sadness, which a trained observer might misinterpret or overlook. This heightened accuracy not only aids in the identification of emotional distress but also facilitates better-targeted therapies. With improved emotion recognition, practitioners can design interventions that specifically address the emotional needs of their patients, leading to more effective treatment outcomes [1, 2].

Furthermore, the wealth of data generated through these analyses enhances research on mental health trends and patterns. By aggregating and analyzing large datasets, researchers can identify correlations between emotional expressions and various mental health conditions, ultimately contributing to a more robust understanding of mental health issues. Thus, the application of deep learning in FER not only promises greater accuracy in emotion recognition but also enriches the overall landscape of mental health research and therapy [3].

Facial Emotion Recognition technology can notably enhance accessibility to mental health services, a vital consideration in today's diverse and often underserved populations. One of the most significant advantages of FER is its ability to facilitate remote monitoring, providing continuous support for individuals who may not have immediate access to mental health professionals. This technology allows for real-time emotional assessment, enabling timely interventions and ongoing support, particularly for those living in rural or underserved urban areas. Moreover, FER technology can bridge gaps for individuals who face barriers to traditional mental health services, whether due to geographical limitations, financial constraints, or stigma associated with seeking help. By leveraging technology, mental health services can reach a broader audience, ensuring that individuals in need receive the support they require.

Additionally, the ability to receive real-time feedback can enhance the effectiveness of mental health treatments. When individuals can monitor their emotional states and receive immediate insights, they are better equipped to engage in their treatment process, fostering a sense of empowerment and agency in managing their mental health. Overall, FER serves as a critical tool in democratizing access to mental health care, ensuring that support is available to all, irrespective of their circumstances [4].

Despite the promising potential of FER technology, significant privacy concerns arise with its implementation. The collection of facial data inherently raises ethical questions about consent and the ownership of personal information. Individuals may not fully understand how their emotional data is being utilized, leading to a breach of trust between users and service providers. Furthermore, the misuse of emotional data can lead to stigmatization or discrimination, particularly in sensitive contexts such as employment or healthcare. For instance, if an employer gains access to an employee's emotional health data, it could unfairly influence decisions regarding promotions or job security, perpetuating biases and discrimination against individuals with mental health challenges.

Additionally, the current lack of rigorous regulation surrounding the use of facial recognition technology poses a significant risk. Without stringent guidelines, unauthorized access to sensitive information could occur, compromising individuals' privacy and safety. Thus, while FER technology offers valuable insights into emotional states, the ethical implications surrounding data privacy must be addressed to ensure that individuals are protected from potential harm.

Another critical consideration is that emotion recognition technology can be inaccurate and misleading. Human emotional expression is inherently variable, influenced by a myriad of factors, including cultural background, individual experiences, and contextual cues. This variability can lead to misinterpretation of emotions, where the technology may incorrectly identify a neutral expression as anger or sadness. Such misinterpretations can have severe consequences, particularly in mental health settings where accuracy is paramount for effective treatment. Moreover, cultural differences in expressing emotions may not be accounted for in the algorithms used in FER technology, leading to biased results that do not reflect the true emotional states of individuals from diverse backgrounds. Over-reliance on this technology may undermine the value of personal human interactions and assessments, which are integral to understanding a person's emotional and psychological well-being. The nuances of human emotion are complex and often require empathetic understanding that technology cannot replicate. Therefore, while FER can be a useful tool, it should not replace traditional methods of emotional assessment and should be used in conjunction with human insight and expertise [5].

Finally, it is essential to recognize that mental health issues are complex and cannot be solely identified through facial expressions. Emotional states are influenced by a multitude of factors, including personal history, environmental stressors, and psychological conditions, which extend far beyond what can be captured through facial cues alone. Relying exclusively on technology for mental health assessments may overlook critical contextual information that provides a more comprehensive understanding of an individual's mental health. For instance, a person may smile apparently while experiencing inner turmoil. Without comprehensive assessments that include personal history and self-reporting, the technology might fail to capture the full scope of their emotional experience. Thus, a holistic approach that integrates both technology and traditional assessment methods is essential for accurate diagnosis and treatment. Comprehensive mental health assessments require an understanding of the individual's background, experiences, and the context of their emotional expressions. Therefore, while FER technology can provide valuable insights, it is crucial to remember that it should complement, rather than replace, a thorough and nuanced understanding of mental health.

1.2 Facial Emotion Recognition and Mental Health

Facial emotion recognition (FER) plays a significant role in identifying the intentions and feelings of humans. FER is used in many sectors, such as to identify whether the person is telling the truth or not in many investigations, to identify autism disorders in adults as well as children, to determine the mental health of an individual, human-computer interactions, warning systems, smart environments, etc. [6, 7]. Accurate determination of facial expressions is important to identify human states and intentions. Various researchers formulate many learning strategies to identify the intentions and feelings of individuals. Among them, convolutional neural networks and recurrent neural networks are identified to provide optimum detection results. Most of the researchers also defined techniques for emotion recognition from facial features and speech signals. However, these features are highly sensitive to the external noises present in the environment. This causes influence in the overall detection results [8].

Accurate identification of the facial expression of an individual discovers the underlying mental health of the individual. Emotion recognition in both children and adults helps in identifying many possible mental health issues as well as helps in diagnosing the externalizing behavior of the individual. These behaviors add influence on the adjustment capabilities of the individual in later life with a higher risk of criminal attitude [9, 10]. The significant facial expressions validated to provide information for behavior identification include happiness, sadness, surprise, disgust, neutral, anger, and fear. Technically, facial expressions are identified based on the functioning of the facial muscles responsible for moving onto different positions for various expressions [11, 12]. The prime requirement of any FER technique to perform well is that the features to be extracted are required to be relevant without any redundant or irrelevant information. Therefore, feature selection (FS) strategies are required to be formulated to avoid redundant information in the image and to improve the classification accuracy with better detection rates [13].

Identification of the optimal features is more important to gain efficient classification output. Most of the FER techniques use the FS phase to select the optimal features to improve the accuracy of classification. Optimization plays a crucial role in efficiently identifying the optimal and most important features that are suitable for classification [14]. Feature extraction (FE) greatly affects detection outcomes; therefore, it is essential to extract suitable features to enhance the technique's performance. The count and dimensions of the features are required to be optimized to acquire better results for FER systems [15]. Extraction of features is an essential part that helps in recognizing the emotions and feelings of individuals. Facial features differ for different emotions and expressions and there are possibly two different approaches available in identifying the features of facial expressions. These two include geometric feature-based and appearance-based approaches [16]. The former approach makes use of geometric parameters to locate facial points such as eyes, nose, etc., and the latter approach uses certain techniques like the Gabor filter, wavelet transform, etc., to extract the features [17]. Feature extraction (FE) converts input data into a collection of new features that provide valuable information for classification, with numerous methods available for this purpose, including Local Binary Pattern (LBP), Gabor filters, and Active Appearance Model (AAM) [18].

Automatically recognizing human facial expressions is gaining attention in diverse domains like surveillance, health care, human-computer interaction systems, etc. [19]. Deep learning and machine learning systems like support vector machine (SVM) and Bayesian classifiers can provide the facility of automatic FER tasks based on the given input facial features [20]. Moreover, automatic FER has many advantages. The labor costs and time required for identification are considerably reduced with the avoidance of human coding. Therefore, it is extended into various fields education, medicine, telecommunication, security, marketing, and automotive industries. Learning models are more efficient in identifying emotions through a thorough learning of the features given as input and the artificial neurons play a vital role in extracting the major features supporting efficient classification [21].

The meta-heuristic algorithms work on efficiently finding the best features from the search space out of all the extracted sets of features. This procedure reduces the depth of the features and effectively lowers the time complexity of the learning process [22]. The choice of a better optimization technique is appreciable for improving the FS phase and reducing the computational cost and dimensionality of features. Meta-heuristics possess the ability to navigate the search space to achieve both local and global optimal solutions, the FS can benefit the researchers in developing better classification results of FER [23].

Emotions of an individual can be identified from various features, including speech and facial features, and these can be extracted not only from images but also from videos. Many researchers have formulated diverse solutions for identifying the emotions of an individual from both images as well as audio and video signals [24]. Efficient classification strategies can improve the outcome of the research with various benefits for different industries, especially the medical field. The psychological stress occurring in children and adults can be deeply understood through emotional recognition [25]. There is even more research concentrating on the area of emotion recognition for finding autism spectrum disorder in both children and adults. Diagnosing and mitigating the defects related to mental health right from childhood is highly necessary to avoid the criminal and antisocial activities of individuals [26].

Technological advancement has led to an increased reliance on image processing and deep learning. Emotions play a crucial role in human interactions and have diverse applications in psychology, behavior studies, and human-computer interaction. Recognizing universal emotions such as disgust, happiness, surprise, anger, fear, sadness, and contempt from facial expressions is a challenging task. Facial emotional analysis involves extracting relevant features from facial actions, such as local and global appearance-based features, as well as deep learning-based features. Automatic expression recognition is complex due to the multi-layered structure of the human face and the range of possible changes in emotions. Human perception is skilled at detecting changes in emotion through continual monitoring.

Recognizing facial emotions is crucial for social interaction, as it provides essential information for understanding the social dynamics between individuals. Deficiencies in facial emotion recognition have been identified in various psychiatric disorders, with presentations of deficits in recognizing emotions such as anger, fear, disgust, sadness, and happiness. Research has shown that individuals with autism spectrum disorder may struggle more in recognizing fear compared to other basic emotions, as indicated in a recent systematic review [27].

Automatic deep learning-based facial emotion recognition could significantly enhance mental health care. It can assist therapists by tracking client emotions, detecting suppressed feelings, and enriching digital therapy environments. This technology may also improve access to mental health services by identifying emotional distress and connecting individuals with specialists. Notably, individuals with conditions like autism spectrum disorders and alexithymia often struggle with recognizing and expressing emotions; emotion recognition tools could help bridge these gaps.

While automatic emotion recognition is unlikely to replace human therapists, it can enable them to concentrate on more complex cases and assist less experienced personnel in managing issues such as burnout or workplace anxiety. Additionally, these tools may aid research, monitoring, and treatment effectiveness. However, implementing emotion recognition technology in mental health care requires addressing ethical, privacy, and safety challenges. Broad societal recognition technologies must establish safeguards concerning privacy, transparency, data usage, and algorithmic biases before being integrated into mental health applications. In this thesis, among many mental health problems, we basically focused on detecting autism spectrum disorder [28].

ASD is a neurodevelopmental disorder characterized by pronounced deficits in social communication, recognizing, processing, and understanding social cues and repetitive patterns of behaviors, activities, or interests. Early intervention benefits children in developing speech and language skills. According to the World Health Organization (WHO), the prevalence of ASD is estimated at one in 59 children.

Some individuals with Autism Spectrum Disorder (ASD) may find it easier to recognize negative emotions like anger and sadness while finding it more difficult to recognize happy faces. However, conflicting findings exist regarding the recognition of happy faces. Moreover, studies have revealed a significant correlation between deficits in facial emotion recognition and reported or objective emotion problems by parents.

One characteristic and defining feature of autism spectrum disorders is the difficulty in recognizing or understanding facial emotions. Due to the core symptoms of ASD, several research studies have been conducted to prove the association of ASD with deficits in facial emotion recognition, and several meta-analyses have confirmed that children and adults with ASD present various deficits in all six emotions of facial emotion recognition. Facial Emotion Recognition (FER) plays an important role in recognizing emotion by identifying facial emotions interpreted as anger, fear, disgust, happiness, sadness, and surprise [29]. Technically, facial emotions are identified based on the functioning of the facial muscles responsible for moving onto different positions for various expressions. Figure 1.1 shows the region of interest in terms of shapes and sizes for facial muscles. It is 127,071 pixels of the total face, where eyebrow size is considered pixels of 27.01%, nose/cheek size is 21.74%, and mouth size is 17.60%.

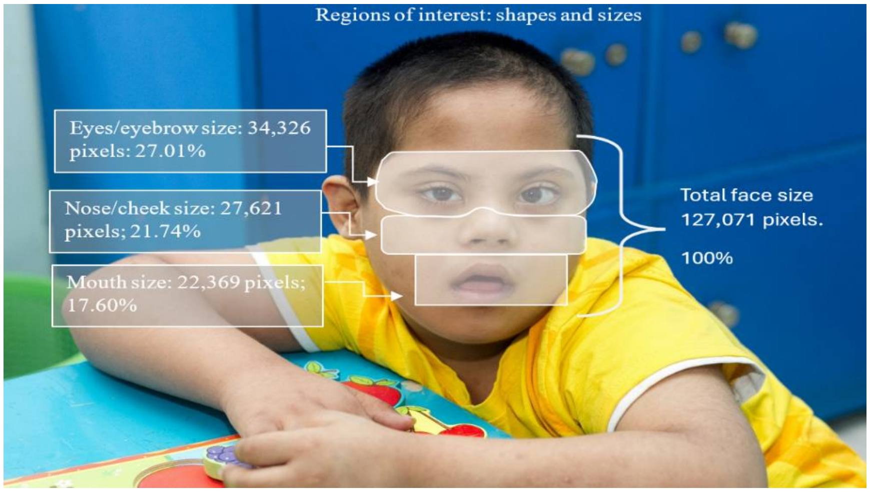

Regions of interest: shapes and sizes
Eyes/eyebrow size: 34,326 pixels: 27.01%
Nose/cheek size: 27,621 pixels; 21.74%
Mouth size: 22,369 pixels; 17.60%
Total face size 127,071 pixels.
100%

Figure 1. 1: Responsible facial muscles.

Any FER technique must perform well because the extracted features must be relevant without redundant or irrelevant information. Therefore, feature selection (FS) strategies must be formulated to avoid redundant information in the image and improve classification accuracy with better detection rates.

1.3 Early Detection and Diagnosis

Challenges related to ASD care stem from the stigma of the condition and the facts of late recognition in both rural and urban areas of society. Consequently, numerous individuals never seek medication or therapy. Moreover, inbuilt biases can contribute to clinician disparities. Datasets do not encompass diversified racial and ethnically marginalized communities, sustaining the disparity. Given that different cultures manifest emotions in various ways, the exclusive acquisition of datasets for a particular mode of emotion expression limits the application of ML models to handling autism spectrum disorder.

Early detection and diagnosis of ASD is essential to provide better therapy and treatment at an early stage. By performing early detection and diagnosis of ASD, these individuals can lead better lives and improve their quality of life. It will also benefit their socioeconomic well-being while reducing their dependence on caregivers and society. The ideal age for ASD diagnosis is between 18 months and 4 years. Early diagnosis allows parents, childhood educators, and caregivers to offer appropriate support and educational intervention per the child's developmental needs. However, children are usually diagnosed between the ages of 4 and 5, and their brains have already passed the critical development window, implying a deficiency in characterizing communicative, emotional, and social signals [30]. Many researchers assert that early intervention may benefit a child's developmental trajectory because their brains are still developing and have more plasticity. Although psychiatric and neurological evaluation is possible, the diagnostic procedure is time-consuming and expensive due to increased healthcare expenses. Instead, machine learning models could efficiently execute the ASD diagnostic procedure by automatically analyzing emotions for early ASD diagnosis and reducing the rise in healthcare costs [31].

1.4 Challenges in Diagnosis and Treatment of Mental Health Issues

Despite advances in research, autism remains a mystery to many professionals. It is a complex brain disorder that affects social interaction, communication, and behavior. Neglecting the exploration of psychiatric and somatic pathologies of autism hinders early prevention of comorbid symptoms. Mental health issues in autism have gained attention, but the evidence is limited. This work aims to explore mental health and other comorbidities in autism populations to support research on prevention and therapy.

Diagnosing mental health conditions in individuals with ASD can be very challenging. For many, it relies on their ability to talk about their thoughts and feelings and express a subjective experience of how the world feels. If the person has great difficulties understanding and using language or cannot use more than a few words, expressing and communicating emotions is difficult. Those individuals may also have a limited theory of mind, which makes them struggle with understanding their feelings and even more with understanding others' feelings. Conclusively, knowing the difference between what an autistic person is going through within the context of autism versus a mental health condition can be challenging [32].

Treatment for people with autism and comorbid mental health problems often remains inadequate. Comorbidity assessment is frequently hampered by the reliance on self-report, which is challenging in people with limited verbal communication. Additionally, the core features of autism may sometimes be hard to differentiate from the symptoms of comorbid psychiatric disorders. It is a common observation in the clinic that some children diagnosed with an autistic disorder lose the core symptomatology of that disorder with time. Still, they continue to have a personality characterized by difficulties in social interaction and narrow interests. These symptoms are similar to those of schizoid personality and Asperger syndrome, and such children may be at particular risk for developing schizophrenia spectrum disorder in late adolescence or young adulthood. Due to the overlap of symptoms between ASD and several psychiatric disorders, clinicians may fail to diagnose a psychiatric disorder in the ASD population until they are well into adulthood [33].

1.5 Research Gaps

As discussed above, the diagnosis of ASD presents a significant challenge due to the variability in symptom onset age and its diverse range of symptoms. Also, the technical facility is unavailable in our country. Moreover, from the literature study, we observed that the lower accuracy had a negative effect on the classification results. Almost all of the available datasets are American and European ethnic based. Furthermore, there are no effective methods of recognizing the facial emotions of ASD with a high degree of accuracy.

So, being motivated, we propose a deep learning-based binary classifier model, which will work as a software. This classifier model will show the results within a few minutes, which proves it can be a good resource for parents and clinical people of Bangladesh. The heterogeneity is evident through varying levels of severity and substantial fluctuations in symptom manifestation. The primary symptom of autism spectrum disorder is an absence of emotional interaction. In the case of impaired affective interaction, understanding facial expressions is critical [34].

Diagnosing ASD is challenging to the naked eye due to several factors contributing to its complexity and variability. These factors include:

a. Heterogeneity of Symptoms
b. Developmental Nature
c. Overlap with Other Conditions
d. Context and Environment
e. Masking and Compensation
f. Subjectivity and Observer Bias
g. Communication Difficulties

So, it requires a multidisciplinary approach that combines thorough clinical assessment, behavioral evaluations, medical history, and input from caregivers and educators. Advanced diagnostic tools, such as standardized assessment instruments and psychological testing, are often employed to provide a more accurate and reliable diagnosis of ASD beyond what can be observed with the naked eye. Currently, there is a lack of effective methods to recognize the facial emotions of individuals with ASD accurately. This limitation in accuracy can negatively impact classification outcomes, potentially causing significant problems. Therefore, having a reliable dataset and a well-designed classifier model is essential for applications based on deep learning in this context [35].

1.6 Aims and Objectives of the Research.

The thesis aims to design several classifier novel models from the available dataset for accurately recognizing facial emotions. Then, using these classifiers, identify individuals with autism spectrum disorder using a high dimensional dataset. The key contributions of this study are listed below:

a. To design an original hybrid Deep Belief Rain Optimization (DBRO) model for emotion classification with better accuracy.

b. To design a novel Bidirectional Elman Neural Network (Bi-ENN) to improve the input information of the training process for accurate classification of FER.

c. To construct a highly efficient Facial Emotion Recognition classifier framework, "InceptionV3DenseNet architecture," using three modalities: audio, video, and text.

d. To create a real-time image dataset, "ASDnet," for special children aged 4-13 for the availability of data dimensions, which is the very 1st time for our country and the whole South Asian region since existing datasets are from America and European ethnic.
[Available link: https://github.com/mashukalamgir/Autism]

e. To architect and validate a novel classification model based on facial Grid-wise expression features for recognizing children with ASD from the ASDnet dataset with improved performance.

Objectives *a*, *b*, and *c* focus on constructing various classifier models for facial emotion recognition with better performance. Objectives *e* and *f* focus on constructing the original real-time Autism individual dataset and ASD detection from the images. On the other hand,

1.7 List of Publications

The proposed research methodologies have generated multiple publications in reputed journals and conferences. These are as follows:

- Alamgir, F.M., Zaman, T., Hassan, M.M., Jonayed, M.R., and Alam, M.S., "**Classification Model for Autism Spectrum Disorder Individuals: Utilizing Facial Grid-Wise**

Emotion Features and Dual-Branch Visual Transformation” *2024 IEEE International Conference on Power, Electrical, Electronics and Industrial Applications (PEEIACON)* 12-13 September, RUET, Rajshahi, Bangladesh.

- Alamgir, F.M., Saif, S.M.H, Hossain, M.S., Hadi, A.A., Alam, M.S., **"Facial Expression Database of Autism Spectrum Disorder Children",** *European Chemical Bulletin*, Volume 12, Special Issue 4, pg. 21109-21120, October 2023. DOI: *doi:10.48047/ecb/2023.12.Si4.1851*
- Alamgir, F.M., Alam, M.S. **"Hybrid multimodal emotion recognition framework based on InceptionV3DenseNet"**, *Multimedia Tools and Applications* (Springer Nature), Volume 10, Issue 23, pg. 1-28, March 2023. DOI: *https://doi.org/10.1007/s11042-023-15066-w*
- Alamgir, F.M., Alam, M.S. **"A Novel Deep Learning-based Bidirectional Elman Neural Network for Facial Emotion Recognition"** *International Journal of Pattern Recognition and Artificial Intelligence* (World Scientific) Vol. 36, No. 10, November 2022. DOI: *https://doi.org/10.1142/S0218001422520164*
- Alamgir, F.M., Alam, M.S. **"An artificial intelligence-driven facial emotion recognition system using hybrid deep belief rain optimization"** *Multimedia Tools and Applications* (Springer Nature) Volume 82, pg. 2437–2464, January 2023. DOI: *https://doi.org/10.1007/s11042-022-13378-x*

1.8 Organization of the Thesis

The thesis can be segmented and analyzed from different viewpoints depending on the reader's focus. The chapters have been arranged based on pedagogical concerns. We start with

Chapter 1 discusses the introduction section with the basics of facial emotion recognition, Mental Health, Autism Spectrum Disorder, and current conditions with expected challenges for Autistic children worldwide, including Bangladesh. Also, the research's motivation, aim, and objective are discussed in depth in this chapter.

Chapter 2 covers background studies and related work on identifying ASD using deep learning models. It also discusses the optimization techniques for the proposed work.

Chapter 3 lays out the steps for making a powered AI system that can recognize facial expressions of emotion using a hybrid deep belief rain optimization method.

In Chapter 4, we discuss the new deep learning-based bidirectional Elman neural network used to recognize facial emotions.

Chapter 5 discussed the Hybrid multimodal emotion recognition framework based on InceptionV3DenseNet.

Chapter 6 covers the ASDnet: Autism Spectrum Disorder detection network Using Grid-Wise Facial Features and Dual-Branch Visual Transformation.

Chapter 7 concludes the work with remaining challenges and recommendations for future work.

Chapter 2: Background Study and Related Works

2.1 Introduction

This chapter elucidates numerous deep-learning models and optimization methods. Since learning a neural network requires adjusting parameters to minimize the loss function, deep learning largely relies on optimization. To guide the network toward the optimal configuration, optimization procedures such as gradient descent are used to update the weights and biases of the network iteratively. Efficient optimization is required for deep learning models to perform faster during training and on average. The latest and most advanced deep learning systems yield the best outcomes when comparing old approaches based on accuracy.

2.2 Understanding Facial Emotion Recognition

Facial expression is one of the primary non-verbal communication methods for humans. More than 50 years ago, Silvan S. Tomkins introduced the concept of basic facial expression. He believed that basic facial expressions were correlated with a few basic emotions. Tomkins' theory significantly impacted psychology, and many researchers arranged his core emotions into four basic emotions: happiness, anger, sadness, and physical shock. Tomkins' theory was later advanced by Carroll Izard, who developed a comprehensive facial action coding system (FACS) for categorizing human emotional expressions [36].

Improvements in facial emotion recognition (FER) techniques allow for various applications in real-world scenarios. These applications range from automotive industries (determining the driver's state) to assistive and telehealth tools. Studies have also shown that FER is effective in supporting mental health for patients and staff, particularly in outpatient and emergency departments. Mental health research often significantly suffers from biased or noisy labels caused by inconsistent annotations across videos and even discrepancies between many human annotators. The annotations come from different instructions that create the potential for inconsistent or extravagant annotations. Additionally, multiple annotators' motivations, moods, and levels of attention can affect their judgments, even for the same expression. Several researchers have illustrated high variance in FER model performances using static or dynamic datasets [37].

2.3 Comparative Analysis of Existing Deep Learning Models for FER

Facial emotion recognition is essential to the software used to identify mental health problems. It has been recognized that FER is often the most influential variable when identifying individuals with mental health problems. Naturally, traditional approaches to FER mainly employ methods to extract features from images. But in recent years, with the development of deep learning, the performance of FER has been further improved. However, the increased complexity of the model has created challenges for real-time FER applications, and to the best of our knowledge, few studies have detailed these challenges [38].

Traditional approaches to FER mainly employ methods to extract features from images of the face and then utilize popular classifiers, such as SVM, to identify the expressions. The performance of traditional approaches depends heavily on the efficiency of feature extraction, making the choice of algorithms and the feature designing process crucial. The most well-known facial expression datasets, JAFFE and CK+, mainly consist of grayscale images of the posed expressions of a few participants. Due to these limitations, traditional FER research usually does not include practical psychological applications and objective FER evaluation. Designing a feature extraction algorithm that performs well on both baseline and practical FER applications is generally difficult. Since the appearance of deep learning methods, when the features were derived from convolutional neural networks through training using massive facial images, the performance of FER has greatly improved. At the same time, the complexity of the model makes the real-time FER application very challenging [39].

2.4 Facial Emotion Recognition for ASD Detection

Facial emotion recognition technology has great potential in ASD diagnosis and treatment, but it is important to approach it cautiously and consider its limitations. Society needs to understand that although facial emotion recognition technology appears to be a remedy for improving ASD diagnosis processes, it has limitations.

The face serves as an information center for understanding emotions, and this tool is the same for the typical population and individuals with ASD. Therefore, facial emotion recognition requires little reliance on understanding context, which people with ASD find very difficult. Detecting emotions from the face is also suitable as an input to demonstrate access to the internal state of individuals with ASD and to teach them "social learning" [40].

In the present century, it was found that there is a significant relationship between the recognition of facial emotions and different mental disorders [41]. It has sparked considerable interest in understanding facial emotion and has been a pivot for developing technology. Facial emotion recognition has found significant relevance in working with cases of autism spectrum disorder, specifically about repetitive behavior, atypical communication, and relationships.

Facial emotion recognition technology is a rapidly growing field in the intersection of artificial intelligence. Researchers in the neurocognitive sciences have long been interested in understanding how humans perceive and understand facial movements and the underlying cognitive and neural mechanisms responsible. This interest in facial expressions has now extended to automatic facial emotion recognition and classification, which is being pursued by computer scientists, especially with the advent of deep learning techniques [42].

2.5 Deep Learning Models for Facial Emotion Recognition

The application of deep learning is found to be the best in identifying facial emotions. Still, most of the available strategies face certain challenges and disadvantages that have major impacts on the results of classification. Certain learning algorithms are slower and require more time to get trained. It reduces the efficiency of the system and increases the computational complexity as well as the computational cost. The techniques highly depend on the input images submitted, but these images consist of several distractions, such as noises due to the environment and poor background. It has a major influence on the classification output. Thus, efficient pre-processing techniques using different improved filters can improve the overall performance and accuracy of the system [43].

The diverse facial expressions of individuals are difficult to identify, and the variations in facial expressions differ based on the individual's age, resolution, etc. Since the face of a human is a 3D rigid object, the minute changes in the expressions are difficult to understand by any technology. It influences the accuracy of the classification system, and the minor changes in expressions, if not predicted properly, might mislead the researchers from producing reliable results. Some of the available techniques are not capable of handling rigid objects, and some are unable to identify the occlusion occurring in the face. Most of the techniques are not capable of providing sufficient and relevant information to the classifier, and certain global features that are provided to the classifier might be environment-sensitive. Many learning techniques use a large count of parameters that increase the computation time of the training process [44].

The introduction of different types of neural networks to compromise the faults of other existing techniques has been improving the FER in recent times. Most of the learning models provided better results in terms of classification, but the training time is high, leading to increased computational complexity and cost. Certain hybrid methods are also introduced to attain higher classification accuracy. The major drawback faced by most of the existing techniques is the lack of essential information to be provided for the classifier, the deficiency of the system to identify the occlusion of the facial images, and the lack of appropriate techniques to remove the environment-sensitive features. All the problems mentioned above stood as a major motivation behind this research proposal. Though there is a huge count of research undertaken to provide a feasible solution, FER is still an area of research to obtain efficient classification results with improved speed and high accuracy [45].

Artificial neural networks can provide an independent solution for extracting facial emotions. Therefore, regardless of variation in facial proportions and non-uniformity of the subjects with ASD, deep learning models for facial emotion recognition from visual data can be used to diagnose ASD using emotional information obtained from facial expressions. In addition, applying facial emotion recognition largely based on deep learning in big data analytics contributes an extra advantage in a practical context in the ASD diagnosis [46].

An increasing number of reports suggest that facial expressions play a major role in diagnosing autism in its early stages. However, analysis of facial expressions remains a challenging problem due to the complexity of emotions and expressions. The deep learning techniques employed for facial emotion recognition offer considerable potential for working towards solving this challenging problem [47].

Facial emotion recognition has become increasingly popular through the development of deep learning models. In particular, numerous deep learning models have been built for better accuracy, feature learning, and improved performance in facial emotion recognition. In this context, detecting autism has also become popular. This condition differs from person to person and makes it difficult to diagnose it in the early stages. As a result, autism is one of the diseases that a person can fight in the later stages of life [48].

2.6 Pertinence of Deep Learning

Deep learning is a sophisticated subset of artificial intelligence (AI) and machine learning. It models its operations after the architecture of the human brain, using artificial neural networks to process large amounts of data. Its remarkable capabilities come from its ability to learn representations of data at multiple levels of abstraction. It effectively mimics human cognitive processes, allowing it to recognize patterns, make decisions, and generate predictive insights.

In some contributions of recent research works, deep learning is tested for automatic emotional image classification and convolutional neural networks (CNN) for facial feature extraction. Some recent research outcomes demonstrate the application of deep learning in ASD detection in different ways, such as pattern discovery, clinical score prediction, correlation contradictions, and real-sense associations. In the disease detection context, ASD was the target of different research works to study and detect using the gold standard of automatic facial emotion and data processing. The most important feature studied is facial emotion preference and its correlation with the relative autistic scores of the measured population. In several studies, automatic systems are applied to the largest open datasets. After deep learning application, the provided main extracted features are correlated based on nonlinear models and represent the interaction between the detected smiling edges [49].

In many research studies, the use of deep learning models is shown to be efficient in capturing emotional facial representations even without handcrafted feature extraction automatically. The faces of children with ASD may possess specific and morphological properties that may be automatically approached by deep learning, facilitating appropriate identification. This way, the study of the facial emotion representing participants with potential ADHD-related symptoms study, as well as autistic traits, is significantly investigated by machine learning techniques. The research on deep learning topics has significantly increased after 2019 [50].

As reviewed above, facial expression recognition (FER) plays a pivotal role in evaluating mental health, particularly within medical contexts. In the later chapters, we presented a handful of innovative approaches to classify seven distinct emotions through a structured methodology encompassing pre-processing, feature extraction, feature selection, and classification. The classification process employs a hybrid technique known as Deep Belief Rain Optimization (DBRO), which demonstrates enhanced performance over traditional methods. Additionally, a bidirectional Elman neural network is utilized for classification, showcasing superior accuracy in predicting emotion labels. Furthermore, the introduction of a Dual-branch CNN-based visual transformation model (Db-CNN-VTM) for identifying autism spectrum disorder (ASD) in children highlights the effectiveness of advanced classification techniques, achieving remarkable performance metrics. Overall, the integration of sophisticated classification methods significantly advances the field of emotion recognition.

2.7 Image Preprocessing Preparations

In this section, we describe several widely used pre-processing techniques to make input images more suitable for processing by deep learning algorithms.

2.7.1 Cropping

Cropping is performed to remove the irrelevant background information and extract the most important part of the image. Cropping is especially useful when the object of interest in the image occupies a small region in a large image. This process registers the original image into a new one that enhances the object of interest. In combination with resizing, cropped images can reduce processing time and reduce memory requirements for learning partitions. For example, when a training partition is derived from a small region cropped from a medical or biological image, resizing may be required to preserve important information about the object and the subject.

2.7.2 Resizing

Before training an image input model, the images are usually preprocessed to conform to the input requirements of a deep learning network, which are typically fixed to a certain size. Resizing is performed to transform an input image into fixed-size training data by preserving most of the important information present in the image. The original aspect ratio of an image should be preserved in the resizing process. Common resizing strategies include normalization (centering), bilateral scaling, and letterboxing. However, these strategies have shortcomings, such as overfitting in certain cases.

2.7.3 Convolutional Neural Networks (CNNs) in Image Processing

This model can deal with the problems arising in spatially organized inputs such as two-dimensional inputs for images. It is also vital to solve the major drawback of traditional neural networks, which is an excessive number of weights. Each unit in the first hidden layer of a deep neural network relates to all units in the input layers, which results in a large number of weights during learning. On the one hand, CNN learns the so-called local receptive field, which restricts connections and depends only on a local portion of the input. On the other hand, it benefits by enlarging the input image. The training weights in CNN are connected adaptively in a hierarchical manner based on the vision science of the human visual system. In the economics of computational save, CNN helps learn feature hierarchies and shows a strong performance specific to object detection problems. More generally, CNN can be used for various image processing with superior performances [51].

Immune cells are important for the human immune system as they can protect the body from disease. It is necessary to perform multiple tasks to protect the human body: killing bacteria or eliminating abnormal cells such as tumor cells in the human body. Various types of cells are involved in the immune system. HIV infects helper T cells, which coordinate an immune response, which leads to a gradual reduction in the immune system. It was also suggested to use the morphology of an immune cell as an indicator for monitoring developing diseases to not obtain the immune cell count from the blood images taken from the patients. It enables one to manage external factors and reduce the costs required. Regarding external factors, we obtained three types of labeled immune cells: inflamed (IC bacteria), non-inflamed (IC normal), and red blood cells, making it difficult to classify only the immune cells that evolved [52].

2.8 Optimization Methods for Deep Learning

Deep learning optimization is essential in training neural networks and improving their performance, convergence speed, and generalization capacities. The model's parameters are adjusted during optimization to minimize a predetermined loss function. Several techniques have been devised to accomplish this effectively. Numerous optimization methods rely on Stochastic Gradient Descent (SGD), which modifies model parameters according to the gradient of the loss function in relation to a randomly selected subset of the training data. An eminent instance is the Adam optimization algorithm, which blends concepts from RMSProp and momentum algorithms. Adam uses squared gradients and moment estimations to adjust the learning rates for every parameter dynamically. Adam's versatility makes him a good fit for various jobs and provides consistency and effectiveness when training. AdaDelta is a notable optimization strategy that adapts the learning rate based on a running average of squared gradients, eliminating the requirement for a manually specified learning rate. AdaDelta is especially helpful when choosing an appropriate learning rate is difficult because of its parameter-free nature. The optimization environment constantly changes with continuous research, and hybrid approaches are becoming increasingly popular. To produce more flexible and potent systems, hybrid learning paradigms blend different optimization techniques or models, combining the advantages of supervised, unsupervised, and reinforcement learning. Optimization strategies are crucial for maximizing the full potential of neural networks in diverse fields like computer vision, natural language processing, autonomous systems, and healthcare, particularly as deep learning applications grow more complicated and varied [53, 54].

Deep learning optimization techniques are essential for training neural networks because they allow efficient parameter adjustments that minimize a predetermined loss function. Numerous methods and algorithms have been created to improve convergence, stability, and generalization of deep learning models. The several optimization methods are as follows:

2.8.1 Stochastic Gradient Descent

An essential optimization technique used in machine learning, specifically for deep neural network training, is called stochastic gradient descent (SGD). Stochastic gradient descent (SGD) is unique in that it selects and processes a tiny portion, or mini-batch, of the data at random for each iteration rather than using the complete dataset to compute gradients. This method works incredibly well with massive datasets because it adds a degree of randomness that helps the algorithm move through the parameter space more quickly. SGD's intrinsic unpredictability is a regularization, reducing overfitting and enhancing the model's generalization ability. The learning rate controls the step size of the optimization process as the algorithm iteratively modifies the model parameters based on the gradients of the loss function. SGD has benefits in efficiency and scalability, but it also has drawbacks, including the need to carefully choose a suitable learning rate and deal with noisy updates. By balancing the stability of batch gradient descent with the efficiency of SGD, variants such as Mini-Batch Gradient Descent improve the methodology even more. Stochastic gradient descent is a mainstay of the optimization toolbox and is essential to the practical training of machine learning models, especially in the challenging field of deep learning [55].

SGD essentially follows the gradient of mini-batches chosen at random downhill. First, a loss function generates the gradient estimate to train an NN using SGD. Subsequently, the iteration k update is applied to the parameters θ. For every mini-batch of m instances in the training set $\{x(1),....,x(m)\}$, the calculations with matching objectives $Y(i)$ are as follows:

$$\hat{g} \leftarrow \frac{1}{m} \nabla_{\theta} \sum_{i} L\left(f\left(x^{(i)};\theta\right), y^{(i)}\right) \tag{2.1}$$

$$\theta \leftarrow \theta - \in_{k\hat{g}} \tag{2.2}$$

The learning rate $\in_k$ is a crucial hyperparameter in this case. Updates become overly dependent on current occurrences if they are too big. If it's tiny, many updates might be required for convergence. Trial and error can be used to determine this hyperparameter. Selecting the learning rate that yields the lowest loss function value is one method. Another method is to utilize a higher learning rate than the optimal one while keeping an eye on the first few epochs. Since the learning rate must steadily decrease over time in practice, it is represented as $\in_k$ iteration k in Equation 2.2.

2.8.2 AdaGrad

AdaGrad is an optimization algorithm that changes the learning rates of model parameters on an individual level. Parameters with high partial derivatives of the loss have an increasing learning rate, while parameters with low partial derivatives have a decreasing learning rate. All of the gradient's squared historical values are used to do this. AdaGrad employs an extra variable r to facilitate gradient accumulation. This algorithm computes the gradient for a minibatch and initializes the gradient accumulation variable to zero at the beginning:

$$g \leftarrow \frac{1}{m} \nabla_\theta \sum_i L\left(f\left(x^{(i)};\theta\right), y^{(i)}\right) \tag{2.3}$$

The squared gradient is accumulated by using this gradient. The update is calculated by modifying the learning rates of all parameters in proportion to the square root of the sum of all historical squared gradients. Lastly, the model parameters are modified with this information:

$$r \leftarrow r + g \otimes g \tag{2.4}$$

$$\Delta\theta \leftarrow -\frac{\varepsilon}{\delta + \sqrt{r}} \otimes g \tag{2.5}$$

$$\theta \leftarrow \theta + \Delta\theta \tag{2.6}$$

Where δ is a tiny numerical stability constant and ε is the global learning rate.

AdaGrad has several drawbacks. It generally works well for straightforward quadratic issues but frequently ends too soon when training neural networks. The algorithm eventually stops before reaching the global optimum since the learning rate is reduced to such an extent. The accumulation of squared gradients during training may significantly decrease the effective learning rate in deep neural networks. AdaGrad works well with some deep learning models, but not all.

2.8.3 AdaDelta

AdaDelta is an adaptive optimization algorithm created, especially for deep learning, to solve problems related to learning rate tweaking in gradient-based optimization techniques. AdaDelta is a learning rate adjustment technique suggested as an enhancement over Adagrad and RMSprop. Its goal is to prevent problems like vanishing or expanding gradients. AdaDelta dynamically adjusts the learning rates for each parameter according to their past gradient data values, unlike typical optimization techniques that call for manual learning rate modification. This flexibility facilitates strong convergence over a range of jobs and systems. Because AdaDelta is parameter-free, it does not require a predetermined learning rate, making it especially helpful when choosing the correct learning rate, which can be difficult. The technique adjusts the model parameters based on a running average of squared gradients. AdaDelta is a valuable addition to the repertoire of optimization strategies in machine learning, having proven beneficial in training deep neural networks due to its self-adjusting mechanism and capacity to handle non-stationary targets.

The primary goal of the AdaDelta method is to address the key limitations of AdaGrad, especially the requirement for a manually chosen global learning rate and a slight decrease in learning rates during training. Instead of keeping past gradients forever, AdaDelta limits their history to a fixed-size window. It stops the building up of squared gradients. AdaGrad starts at the start of training and adds up the squared gradients from each round, as explained in the previous section. During training, this cumulative total keeps increasing, reducing the learning rate across all dimensions. The learning rate eventually gets infinitesimally small after many cycles. Rather than accumulating to infinity, AdaGrad becomes a local estimate using current gradients with the windowed accumulation. Thus, even after numerous updates have been made, learning keeps progressing [56].

2.8.4 RMSProp

Neural network training frequently uses the Root Mean Square Propagation (RMSProp) optimization algorithm. Adding a system that modifies the rates for each parameter separately during the optimization process tackles issues related to learning rate adaptation. RMSProp adjusts the learning rate inversely proportionate to the square root of the moving average it keeps track of, which is a moving average of the squared gradients for every weight. The approach is handy when features have varied sizes or when working with non-stationary targets because of its adaptive scaling, which dampens learning rate oscillations. RMSProp is known for converging fast and effectively on various deep-learning tasks. Its variable learning rate aids in stabilizing and improving neural network training by reducing the effects of exploding or disappearing gradients. Despite being well-liked and frequently utilized, RMSProp is frequently combined with other optimization algorithms, like Adam, to take advantage of the benefits of diverse strategies for enhanced performance in a range of machine learning applications [57].

RMSProp is an additional algorithm that alters AdaGrad. An exponentially weighted moving average should be used instead of the gradient accumulation in order to improve the performance in the nonconvex case. AdaGrad reduces the learning rate by using the complete squared gradient history. Instead, RMSProp discards history from the extreme past using an exponentially decaying average to quickly converge upon locating a convex bowl [58].

2.8.5 Adam

Adaptive Moment Estimation, or Adam for short, is a well-liked optimization technique frequently used in deep neural network training. Adam is an optimization technique that integrates the advantages of both momentum and RMSProp techniques, offering a more effective and flexible approach. It keeps two moving averages: one for the squared gradients (type of RMSProp) and another for the gradient of the loss function (comparable to momentum). By computing adaptive learning rates for each parameter using these moving averages, Adam can dynamically modify the step size depending on the gradient information from the past. Adam also uses bias correction to offset the impact of the moving averages in the early training phases. With the help of its momentum term and adaptive learning rate mechanism, Adam can handle sparse gradients well, converge rapidly, and maneuver through challenging loss landscapes. Adam has succeeded in several deep learning tasks, although it might not always beat alternative optimization techniques in every situation. For best results, hyperparameters should be carefully considered. Despite this, Adam continues to be a popular and adaptable optimization technique that significantly enhances the performance of deep learning models in various applications [59].

Adam is a popular optimization method in deep learning. The term originates from adaptive moment estimation, which provides varying adaptive learning rates for distinct parameters by analyzing the first and second derivatives of the gradients. Adam integrates the advantages of RMSProp in varying conditions with those of AdaGrad in situations characterized by sparse gradients.

2.8.6 Deep Learning Optimization Techniques in Facial Recognition

Facial recognition is an essential tool in processing facial information that includes computers recognizing and measuring a person's face in a digital image. Applications in pattern recognition, identity verification, authentication, computer-human interface, automated video surveillance, electronic commerce, health, finance, and other fields have been widely utilized. The two predominant facial recognition technologies are identity and verification. Face identification is the process of defining a person's identity using their facial photographs. The system must indicate whether the estimate is accurate or incorrect during face verification, given the face image and identity estimation. Despite the widespread use of face recognition technology in systems and applications, finding the best ways to achieve high accuracy and low computing overhead remains a complicated problem. Even after years of work, many academics still face obstacles in their quest to understand facial recognition fully. Several techniques, including Eigenface and Fisherman's approaches, have been used up to this point. These methods use algorithms like Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Independent Component Analysis (ICA), and similar techniques. Face recognition software typically employs the Support Vector Machine (SVM) classification technique and the Local Binary Pattern (LBP) methodology [60].

Furthermore, to solve the difficulties in facial identification, neural network approaches, sometimes called deep learning approaches, including the convolutional neural network (CNN), Furthermore, to solve the difficulties in facial identification, neural network approaches, Deep learning techniques, such as the convolutional neural network (CNN), have been used. Deep learning is a recent and promising study area in computer vision and machine learning, especially in image recognition and dimensional reduction. Deep learning is a form of machine learning that uses multi-layer perceptrons (MLPs) that include numerous hidden layers and adjusts neural network structure. Deep learning uses a model architecture of several nonlinear transformations to extract high-level characteristics from the input. This feature creates a hierarchical representation by drawing on the lowest level [61].

Compared to the previous method, the newly developed deep learning algorithms' accuracy yields the best results. It is anticipated that face recognition technology will find greater use due to technological advancements, which motivates academics to continue their investigation into optimizing face recognition methods. Developing on this concept, researchers can integrate current techniques, such as merging traditional methods with deep learning or traditional methods with simply traditional methods, to provide optimal solutions.

Image dimension processing and identification have been effectively implemented through a new area of research known as deep learning in computer vision and machine learning. Deep learning achieves high accuracy using an Artificial Neural Network with a multi-layer perceptron (MLP). Deep learning is beneficial for addressing data-rich problems through a multi-layered architecture involving multiple nonlinear transformations. A study shows that deep learning architecture outperforms conventional techniques in contemporary situations requiring complicated duties like computer vision and natural language understanding. The study also noted that deep learning might use multilayer architectures to handle complicated problems, which shortens problem-solving times and improves accuracy. The multilayer architecture in deep learning is an application of the subsampling technique. Because of this, deep learning is incredibly effective at resolving complicated issues [62].

2.9 Related Research Works

During this thesis writing, numerous research papers were thoroughly reviewed. Below are highlighted the most relevant and updated papers.

Salama et al. [63] introduced a multimodal emotion recognition system with the help of 3D-CNN architecture. Three different types of emotion recognition systems were built such as the EEG-based, face-based, and fusion-based approaches. The technique used the Spatio-temporal features extracted from the electroencephalogram (EEG) signals and the video extract of human faces to identify the different types of emotions. Data augmentation was carried out and the ensemble learning technique was put forth to obtain the final fusion-based prediction results. The 3D-CNN was used for the EEG-based approach to carry out the prediction and the SVM was used by the face-based approach to obtain the classification results. The fusion was carried out for the multi-modalities based on the bagging and stacking methods. Experiments carried out proved that the technique was able to classify emotions with better accuracy. Among all the techniques, stacking was seen to provide better accuracies of 96.79% for arousal and 96.13% for valence classes, respectively.

Hai–Duong Nguyen et al. [64] presented an architecture of CNN known as the multi-level CNN (MLCNN) to classify the emotions of human faces. The MLCNN was able to identify the major mid-level and high-level features from the dataset, and the irrelevant information was effectively eliminated to improve the classification accuracy. The MLCNN consists of an 18-layer CNN where the multi-level features were utilized to train the model. The features from the 2nd, 3rd, and 4th layers were used, and the information from the 1st layer was retarded as it covers trivial information like hair, background, etc. Moreover, the technique employed a 256-unit fully connected layer to retard the irrelevant information and to identify only the information related to the facial expression. The relevant characteristics were combined, and the classification was executed. The simulations showed that the scheme outperformed other existing techniques, achieving an accuracy of 74.09%.

Wang Xiaohua et al. [65] presented a framework called the two-level attention with two-stage multi-task learning (2Att-2Mt) to identify and classify facial expressions on static images. Initially, the images were utilized to extract the features and then were enhanced with the employment of a bi-directional recurrent neural network (Bi-RNN) with second-level attention. Then, the prediction was carried out with the use of a two-stage multi-task learning architecture. The loss function was minimized with the use of Turkey's bi-weight loss function to improve the accuracy, and hence, the influence caused by the trivial samples was eliminated. The experiments carried out proved that the technique was able to provide better prediction results with better accuracy.

A novel technique was introduced by Luefeng Chen et al. [66] for identifying facial expressions in human-robot interaction. The technique used the deep sparse autoencoder network (DSAN) to regulate the learning process of the facial features with the consideration of the sparsity of hidden layers in learning algorithms. The SoftMax regression was used for classification purposes. The technique called SoftMax-regression-based deep-stacked autoencoder network (SRDSAN) was used to identify the facial expressions of humans during the interactions exactly. The technique efficiently faced the problem of local extrema and the gradient diffusion problems were effectively optimized through the tuning of weight parameters in the training process. That process improved the robustness of the system with improved performance efficiency in computing the facial expressions of humans.

A contemporary scheme was proposed by Alia K. Hassan et. Al. [67] to identify facial expressions through graph mining. Both the graph theory and mining were used in the technique to identify the sub-graphs of every emotion class with the use of the gSpan algorithm. The graphs were generated based on the face region into nodes and edges. The overlap ratio metric was used to decrease the number of generated sub-graphs. Finally, the binary classification was carried out to classify the input face image into six different levels of classification. The graph theory helped the system to reduce the overall complexity while classifying the facial images. The technique showed better results in simulations, achieving a 90% accuracy rate for the SAVEE dataset, exceeding most classification methods.

Arpita Gupta et al. [68] suggested a new technique to recognize the facial emotions of an individual based on deep learning. The technique used the deep residual network along with inductive learning and attention module for efficient classification of images without labels. The self-attention module was chosen to improve the visual perspective of the model. The major aim of the technique was to overcome the problem of dealing with the FER image datasets without labels. Based on the inductive learning and attention module, the technique used transfer learning to identify facial expressions. The simulation results revealed that the technique performed well in identifying the facial expressions with reduced over-fitting and improved accuracy. The results produced 56.21% and 57.8% for 50 layers and 152 layers, respectively.

Dazhi Jiang et al. [69] introduced a technique to solve high-level human emotion recognition problems. Initially, the feature extraction was carried out where 16 visual features and 14 audio features were extracted. These features were used in the technique to train the audio and video classifiers. The lyrics contained in Motion Emotional Unit (MEU) were saved as positive and negative using the K-nearest neighbor algorithm, which acted as the lyric's classifier. Finally, all three classifiers were combined using the classification strategy called probability and integrated learning (PIL), where the emotions were classified. The classification algorithm was able to adapt the classification fuzziness created due to uncertainty. Three diverse analysis methods were presented such as the emotional tube, emotional decision preference, and emotional sensitivity based on the classification probability. The experimental results illustrated that the performance of the approach was more optimal than the other approaches for music videos.

Identifying human emotions from EEG signals based on a deep learning scheme was presented by Rahul Sharma et al. [70]. Initially, the features were decomposed using the DWT transform into rhythms, and the higher-order statistics were used to identify the non-linear dynamics of the rhythms. For this purpose, the third-order cumulants (ToC) were chosen. After decomposing the signals, Particle Swarm Optimization (PSO) was used to constrain the dimensionality of the data to a specified threshold. The classification was executed using the long short-term memory (LSTM) neural network, wherein the ToC coefficients were categorized through the softmax layer. The simulation results showed that the technique outperformed the other methods due to the implemented reduction strategies. The simulations produced 82.01% classification accuracy.

Martina Rescigno et al. [71] introduced a personalized model to identify facial emotions in valence/arousal dimensions. The technique used transfer learning for classification purposes, where the information learned by the CNN was exploited to produce a subject-specific model. The major relevant features were extracted to boost the classification results. The experimental results suggested that the technique proved to work well in identifying the facial emotions in valence/arousal dimensions. The model also provided a very low error rate, and the RMSE values are 0.09 for valence and 0.1 for arousal, respectively.

Xiao Liu et al. [72] presented a FER technique using the extracted facial geometric features. The geometric features were extracted from the landmarks that were identified based on the movement of the facial muscles. The pixel location of the landmarks was described based on the reference frame. The subset of landmarks was described, and the geometric center relevant to all the landmarks was identified. After the extraction of relevant features, a combined model of classification was introduced to identify the facial expressions. The technique used the SVM classifier combined with the genetic algorithm (GA) for multi-attribute optimization with parameter and feature selection. Finally, the experiments were carried out to analyze the efficiency of the technique where the multimedia understanding group and extended Cohn-Kanade dataset were used. The results provided better performance compared to other approaches.

Gozde Yolcu et al. [73] invented a deep-learning strategy to identify the facial expressions of customers. The customer behaviors were monitored to identify the interest shown in various items. Initially, the customer's attention was recognized using the head pose estimation. The head orientation of the customers based on their interest in certain advertisements was noted to describe the interests of every customer. Meanwhile, facial expressions were also noted to identify the interest shown by them. The frontal faces were first identified, and the most relevant facial expressions were segmented, and then the iconized face image was generated. The confidence values were computed for the iconized face image, and the expressions were analyzed. The experimental results showed that that system provided better robustness.

Zhongke Gao et al. [74] introduced an EEG-based emotion recognition technology to analyze the emotions of humans based on an optimization approach. The binary coding system and the GPSO were combined to produce the automatic optimization framework that was capable of searching the search space more efficiently to allow the CNN to generate efficient and accurate prediction results. The optimized CNN was applied for the emotion recognition task over EEG signals, and the efficiency was identified through the evaluation of the model over the movie-evoked emotion recognition task. The results suggested that the model could produce efficient classification outputs.

Di Wu et al. [75] presented an emotion recognition model for the EEG signal using the long-short-term memory (LSTM) neural network. The facial videos and the EEG signals of the persons watching emotion-stimulated videos were collected and produced as input to the model. The important facial features and the EEG features were extracted based on the fully connected network at every point. The correlation of the LSTM was explained through the application of the self-attention mechanism at varied hierarchies. Human emotions were recognized with the help of selective focus, which improved the utilization of EEG signals. The key signal frame for every time point was identified using the temporal attention mechanism. The experimental results showed the technique could accurately recognize human facial expressions.

Quang Tran Ngoc et al. [76] introduced a facial landmark emotion recognition technique based on deep learning. The technique made use of the directed graph CNN (DGCNN) that worked with the use of facial landmark features as information for classification. The graph edges were constructed based on the Delaunay method and the nodes were described with the facial landmarks. The graph neural network helped to identify facial emotions based on the geometric and temporal information of the face. A stable temporal block was used in the graph neural network to overcome the vanishing gradient problem. The experiments carried out suggested that the technique was able to classify the emotions of the human faces more efficiently than the other compared techniques.

Ilyes Bendjoudi et al. [77] suggested a technique for context-based emotion based on the multi-label multi-task CNN technique. There were three prime phases in the technique such as a body feature extraction module which was a pre-trained Xception network, a scene feature extraction module centered on the VGG 16 network, and a fusion decision module. The major objective of the technique was to deal with the unbalanced emotion classes. The training was carried out, and the MFL loss function was compared with two other loss functions to understand the behavior. Upon simulations, the MFL scheme was found to outperform the other Euclidean and binary cross-entropy loss functions.

Tsangouri et al. [78] proposed EmoTrain, a deep-learning communication platform to improve facial expression recognition and reciprocity skills in children with ASD. EmoTrain is a training platform that enables users to relate facial emotions pictures displayed with real faces in real-time. The efficiency of this method is assessed by administering training with EmoTrain to a set of members affected by ASD. The data collected from member performance via EmoTrain is analyzed, and their face-processing skills are compared after and before playing to establish an effective platform.

Manfredonia et al. [79] for ASD, facial expression is reduced. The capability of individuals with ASD to make facial expressions of emotions in response to verbal prompts. Differences in the ability to express specific emotions in response to the prompt were linked to parental descriptions of communal relation skills. It specifies the possibility of face reactions as the target for involvement and uses automated facial expression identification devices as analytic and result measures in ASD.

McIntosh et al. [80] examined the automatic facial expressions of adults and children with ASD and a control group matched on gender, verbal intelligence, and gender. Members observed images of angry and happy expressions, though the movement of their brow muscles and cheek parts was monitored with electromyography. ASD children cannot mimic facial expressions automatically, but normal members do. The information proposed that autism is related to impairment of the simple automatic social communication process. Results suggest understanding usual and unusual social awareness.

Gordon et al. [81] proposed the dataset of children with ASD and IQ matched; typically developing children were skilled at making angry and happy expressions with the FaceMaze computer game. This FaceMaze is a cost-effective and good training program in facial expression production, which is attractive for children and has a familiar setting and is safe. After and before playing the happy and angry reactions of FaceMaze, kids posed angry and happy expressions. The results show the connection between disgust and angry facial expressions by indicating the facial expression quality of disgust in angry facial reactions.

Tang et al. [82] proposed the conventional neural networks (CNN) based technique for facial expression for children in social environments. Understanding children's facial expressions has absorbed more attention in pediatric medicine and psychology, which is ASD. The novel dataset, the RCLA and NBH smile dataset, is presented. Initially verify the validity of our approach on two novel datasets, and then we train the method on our child dataset and absorb suitable results. The results show that it is possible to find a child smiling in an actual environment automatically.

2.10 Summary

The problem of detecting mental health problems from facial emotion is not trivial as it includes extracting features, reducing dimensionality, handling a large-scale problem due to multi-class or multi-label, predicting accurately, choosing the right training/testing approach, class embedding, and so on. Researchers are encouraged to look further into optimizing the face recognition approach since it is anticipated that as technology advances, face recognition technology will find more significant applications. Proceeding from this concept, researchers can integrate current techniques, including merging conventional methods with other traditional methods or merging traditional methods with deep learning, to attain optimum outcomes. We believe that with further advancements in emotion recognition, we can extract or predict many more relevant outcomes that could be tackled using the latest techniques in this domain.

Chapter 3: Intelligent Facial Emotion Recognition Construction

3.1 Introduction

From the survey conducted, it was found that there is no proper approach to identify the facial emotions of people with a better accuracy level. Most of the methods primarily adopted learning algorithms to detect facial expression, where few based their approaches on the meta heuristics principle to identify the emotions. Well, the existing methods for facial emotion recognition even do not support the required accuracy, as lower accuracy may deteriorate the results of the classification, and this, in turn, will give rise to some crucial problems, especially in the medical area. This chapter focuses on monitoring individuals' facial expressions to detect them with greater accuracy. The precise detection of varied facial emotions depends heavily on the use of geometric and appearance-based elements. The Deep Belief Network (DBN) is used for effective classification. For weight optimization, the Rain Optimization Algorithm (ROA) is used. The objective of the proposed framework described in this chapter is to enhance the research domain of facial emotion identification by achieving precise classification outcomes.

The primary contributions of the proposed chapter are as follows:

a) This chapter proposes a novel hybrid approach Deep Belief Rain Optimization (DBRO) classification model for recognition of diverse facial emotions of human beings. This novel method improves classification accuracy by merging Deep Belief Network (DBN) and Rain Optimization Algorithm.

b) A unique Multi-Objective Seagull Optimization Algorithm (MOSOA) is proposed to reduce feature dimensionality and select the most important features for classification.

c) The evaluation of the performance of the proposed approach in terms of performance metrics was weighed against the other facial emotion recognition strategies.

3.2 Proposed Method

This chapter introduces a new and innovative technique called Deep Belief Rain Optimization (DBRO) for accurately classifying facial emotions. Preprocessing, feature extraction, selection, and classification comprise the method. Histogram of Oriented Gradients (HOG) and Gabor filters extract geometric and appearance-based image characteristics after noise removal. Multi-Objective Seagull Optimization Algorithm selects category-specific features. Subsequently, the DBRO approach is used to classify facial expressions. The framework operates in automatically identifying the facial expressions of individuals, which would result in better classification with a higher accuracy level. Figure 3.1 shows the global framework.

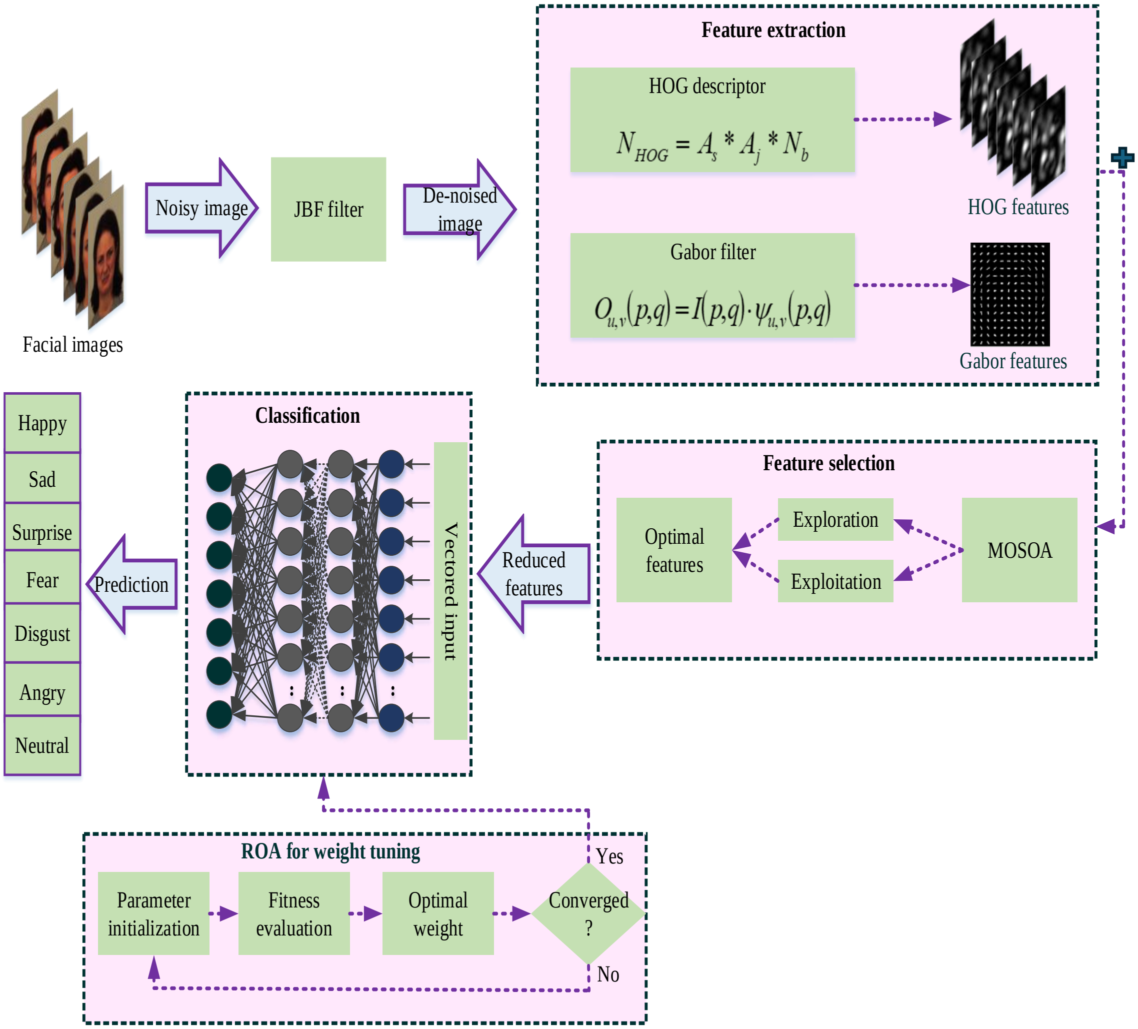


Figure 3. 1: Global architecture of proposed emotion classification framework.

3.2.1 Image Pre-processing

The purpose of this module is to enhance the quality of images by minimizing the presence of noise and distortions using the joint bilateral filter (JBF) technology. This filter efficiently maintains edge information in images, compared to traditional bilateral filters that only preserve edge information based on image intensities. The core primary component is used as the reference image, eliminating the range-filtering kernel to address this limitation.

The pre-processing technique's structural diagram is depicted in Figure 3.2, illustrating its efficiency in preserving edge information.

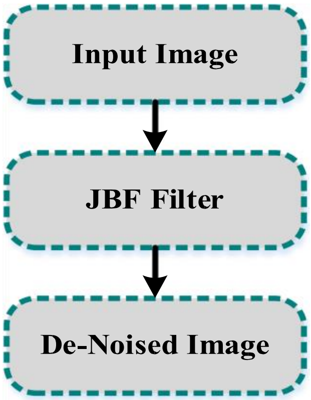


Figure 3. 2: Structural Diagram of Preprocessing.

3.2.2 Feature Extraction

The HOG and Gabor filter techniques are used to pull out the most important geometric and appearance-based features after the pre-processing stage. HOG captures the distribution of intensity gradients or edge directions, while the Gabor filter, a linear filter used for texture analysis, effectively captures frequency and orientation representations of the image.

(i) HOG for Geometric Feature Extraction

The Histogram of Oriented Gradients (HOG) is a technique used to extract features from images, specifically for the purpose of recognizing facial emotions by identifying the main facial components. It remains unchanged by changes in lighting and shape and effectively captures distinctive characteristics. The suggested method uses the HOG descriptor to identify prominent facial features like the eyes, nose, and mouth. Geometric points are placed on these components to derive geometric characteristics. The Histogram of Oriented Gradients (HOG) is calculated using the orientation and gradient data obtained from the image's first derivative. The HOG descriptor is used to extract gradient and orientation information from images that are partitioned into cells and overlapping blocks. The histogram bins contain stored cell orientation information, which is then sorted and organized into a histogram. The technique guarantees precise and effective picture processing.

(ii) GF for Appearance-based Feature Extraction:

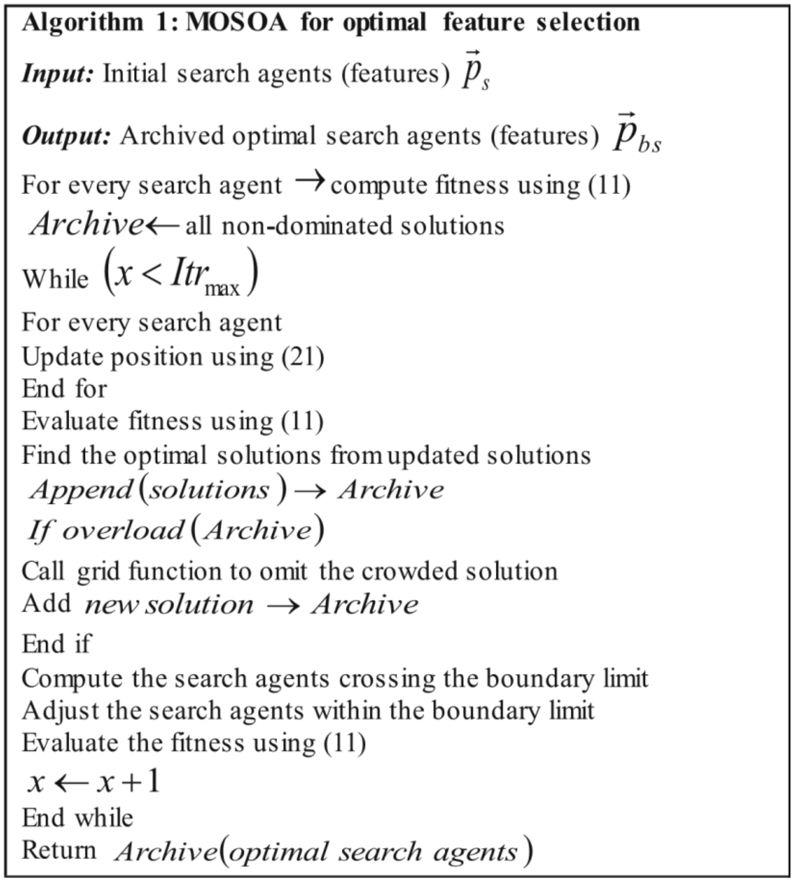
The Gabor filter (GF) is a highly favored technique for extracting features because of its ability to localize spatial frequencies and its resilience to variations in illumination and picture noise. It remains unchanged regardless of changes in scale, rotation, or transition. The proposed study uses the GF to extract appearance-based elements from facial picture samples, which are essential for detecting various sorts of emotions.

After both the features have been extracted from approach, they are added together and forwarded to the next step to make the overall classification better.

3.2.3 Feature Selection

This module focuses on feature selection for accurate classification, aiming to eliminate unwanted features and select only necessary ones. A Multi-Objective Seagull Optimization Algorithm (MOSOA) has been proposed to automatically select relevant extracted features for prediction. The MOSOA algorithm mimics the complex actions of seagulls throughout their migration and hunting activities, including the process of exploring and exploiting their surroundings in search of prey. Seagulls engage in group migration, strategically positioning themselves to prevent collisions, following the most proficient seagull during exploration, and adjusting their positions dependent on the leading seagull's location. These characteristics are used to construct a multi-objective optimization problem for the purpose of choosing relevant features. The MOSOA can be applied to other bird migrations and can be used to optimize the selection of features for accurate classification.

The search agent's position is computed using a variable to avoid collisions while exploring the search space. Those search agents who successfully avoided the collision proceeded toward their nearest neighbor. The position of the search agent is chosen based on the optimal group search agent. MOSOA optimizes search agents. It analyzes search agent fitness and updates the best-archived solutions. If archive capacity is reached, the grid approach is used to eliminate excessively crowded solutions. Then, the newly found solution is included in the archive, and the bounds of the search agent are modified. The optimal search agent receives the new position. Search agent aggression drives exploitation, using a method of evaluating the best Pareto optimum solution in relation to the current solutions. A leader is selected for the purpose of accomplishing this goal. The archive uses the roulette wheel selection method to find the least crowded space with the best solutions inside the optimal border. The archive stores the most efficient search agents or optimal characteristics and excludes irrelevant ones.



Algorithm 1: MOSOA for optimal feature selection

```
Input: Initial search agents (features) $\vec{p}_s$
Output: Archived optimal search agents (features) $\vec{p}_{bs}$
For every search agent → compute fitness using (11)
$Archive \leftarrow$ all non-dominated solutions
While $(x < Itr_{\max})$
    For every search agent
        Update position using (21)
    End for
    Evaluate fitness using (11)
    Find the optimal solutions from updated solutions
    $Append(solutions) \rightarrow Archive$
    $If\ overload(Archive)$
        Call grid function to omit the crowded solution
        Add $new\ solution \rightarrow Archive$
    End if
    Compute the search agents crossing the boundary limit
    Adjust the search agents within the boundary limit
    Evaluate the fitness using (11)
    $x \leftarrow x+1$
End while
Return $Archive(optimal\ search\ agents)$
```

The feature selection phase focuses on selecting optimal features to accurately describe individual emotions while also reducing dimensionality issues to achieve maximum accuracy rate by focusing on the necessary features, thereby reducing the overall complexity of the classifier.

3.2.4 Classification

The objective of this module is to divide photos of people into seven different groups based on the expressions on their faces. It will be achieved by employing a hybrid Deep Belief Rain Optimization technique. The technique assigns labels to images based on input attributes. The study includes specimens from the KDEF and JAFFE dataset.

The suggested DBRO classification method uses Deep Belief Networks (DBN) for the training and classification of images according to expressions. The Deep Belief Network (DBN) consists of unsupervised architectures of Restricted Boltzmann Machines (RBMs) structured into input, hidden, and output layers. The hidden layers are crucial for feature detection, while the classifier is trained using selected features to assign appropriate class labels. The model ensures mutual independence among layer units for efficient training, utilizing a probabilistic approach with random weights and biases. The fine-tuning of the classification model's parameters is achieved through the ROA algorithm, facilitating accurate labeling of testing samples based on the learned features.

3.2.4.1 Dataset Description

The proposed model is used to figure out how people are feeling by looking at their faces. Face features change a lot when people are feeling different emotions. The classifier helps in discerning differences among emotions by labeling the features accordingly. The KDEF dataset and the JAFFE dataset, which include facial images displaying various emotions, were selected for examination [83]. The KDEF dataset contains 4900 images of men and women with various facial expressions. The proposed research uses 970 images for training and 194 for testing.

The JAFFE dataset is a collection of images used for facial emotion recognition. It includes data from 10 female Japanese participants and 213 image samples. The dataset contains pictures with unique pixel resolutions, each representing one of seven facial emotions: happy, sad, fearful, surprised, disgusted, neutral, and angry. It also contains training, testing, and validation samples.

The JAFFE database has been created from openly available information and contains 213 photos of facial expressions from 10 Japanese females. Each participant displayed six primary emotions along with a neutral expression. The distribution of feelings was as follows: 30 instances of anger, 29 of disgust, 33 of fear, 30 of happiness, 31 of sadness, 30 of surprise, and 30 neutral emotions. Each emotional expression consisted of three to four images. The image is in grayscale. All facial photos were captured under strictly controlled conditions with identical lighting and no occlusions like hair or glasses.

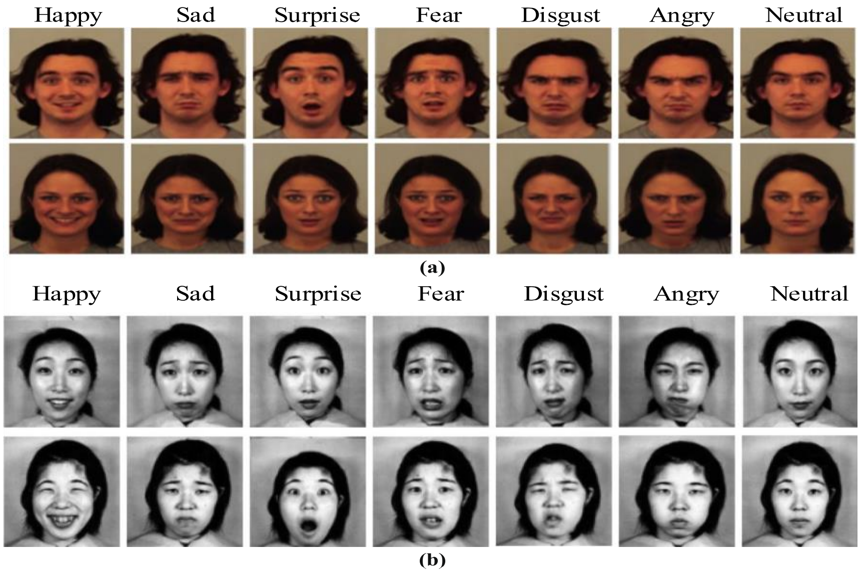


Figure 3. 3: Facial expression samples from (a) KDEF and (b) JAFFE datasets.

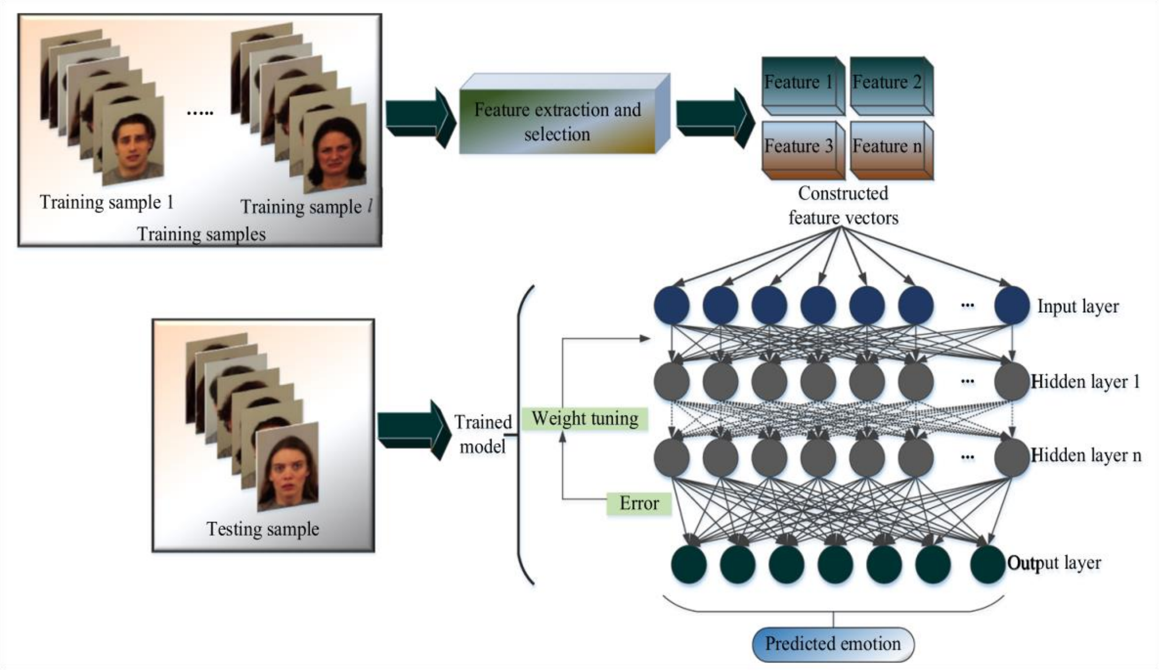


Figure 3. 4: Training and testing procedure of the Deep Belief Rain Optimization (DBRO) methodology.

Figure 3.4 shows the training and testing process of the proposed Deep Belief Rain Optimization (DBRO) method, which is inspired from Deep Belief Network(DBN) and Rain Optimization Algorithm(ROA). The DBRO classifier is trained with a data set of photographs of people with all sorts of facial expressions. Unsupervised Restricted Boltzmann Machine (RBM) develops DBN network models with input, hidden, and output layers. The features are identified based on individual input in these hidden layers. The DBN model is trained to give correct class value predictions using discriminative input features. This DBRO approach makes each layer to be mutually independent all the time.

Further, it is a probabilistic method with random weight and bias. The model is further optimized through the Rain Optimization Algorithm for every class through training. Since we blended the Rain Optimization Algorithm and Deep Belief Network, hence we called this method as Deep Belief Rain Optimization (DBRO) method.

Testing samples are divided based on the refined model. The DBN model is made up of several probabilistic models and the output layer. The first hidden layer then becomes the subsequent visible layer. Training is sequential, with each RBM trained in turn. It allows the identification of images based on the input features presented. The Rain Optimization Algorithm (ROA) alters the weights to make DBN perform better. This is by changing the value of the weight parameter such that it gives an optimal value for the neural network. This approach allocates DBN weights in analyzing search agent fitness in solution space. The weight radius sets the maximum weight parameter since larger radius entails higher weight values whereas smaller radius begets lower weight values. The evaluation of the limit of the weight parameter to select the optimal weight from all available options. The test involved reconstructing features for all classes, computing reconstruction losses, and optimizing errors so that the classifier can label them into different emotions.

The input features for learning are shown in the visible layer, while the output layer masks these input features to resemble the original submitted input, which is later restored. The joint probability of the visible and hidden layers can be represented as:

$$\mathrm{P}(\nu, h) = \frac{e^{-\mathrm{E}(\nu,h)}}{\int\int_{\nu,h} e^{-\mathrm{E}(\nu,h)}} \tag{3.1}$$

where $e^{-E(v,\ h)}$ denotes the energy function. The Gaussian energy function is superior to the binary energy function and can be expressed as follows:

$$\mathrm{E}(\nu, h) = \sum_j \frac{(\upsilon_j - a_j)}{2\sigma_j^2} + \sum_k \frac{(h_k - b_k)^2}{2\sigma_k^2} - \sum_{j,k} \frac{\upsilon_j}{\sigma_j} \frac{h_k}{\sigma_k} \omega_{jk} \tag{3.2}$$

where, $\upsilon_j \ and \ h_k$ denotes the activation states of visible and latent layer units $j \, and \, k$, b_k denotes the bias values, $\sigma_j \, and \, \sigma_k$ are the standard deviations for the input and hidden units with a value around 1 and ω_{jk} denotes the weight parameter connecting $\upsilon_j \ and \ h_k$.

The conditional probabilities of the Restricted Boltzmann Machine can be derived using Bayesian inference as follows:

$$\mathrm{P}(h / \nu) = \frac{\mathrm{P}(\nu, h)}{\mathrm{P}(\nu)} = \frac{e^{-\mathrm{E}(\nu,h)}}{\int_h e^{-\mathrm{E}(\nu,h)}} \tag{3.3}$$

$$\mathrm{P}(\nu / h) = \frac{\mathrm{P}(\nu, h)}{\mathrm{P}(h)} = \frac{e^{-\mathrm{E}(\nu,h)}}{\int_\nu e^{-\mathrm{E}(\nu,h)}} \tag{3.4}$$

The hidden units in the RBM can be adjusted using the specified visible unit through the normal distribution as follows:

$$\mathrm{P}(h_k \mid \nu) \approx N(\lambda_k, \sigma_k); \qquad \lambda_k = b_k + \sigma_k \sum \frac{\upsilon_j}{\sigma_j} \omega_{jk} \tag{3.5}$$

$$\mathrm{P}(\upsilon_j \mid h) \approx N(\lambda_j, h_j); \qquad \lambda_j = a_j + \sigma_j \sum \frac{h_k}{\sigma_k} \omega_{jk} \tag{3.6}$$

RBMs are stacked on top of each other in the DBN model, which ends with an output layer. Once the first hidden layer is trained, it becomes the visible layer for the next hidden layer. This continues until the last RBM is trained. A polling classifier is added to the last hidden layer after the pre-training process to make the labels for the images. The model is trained with the input characteristics in the following manner, and the images are labeled correctly based on the input features. These parameters can be used to describe the whole model:

$$\delta_{DBRO} = \{\delta_{\omega}, \delta_b\} \tag{3.7}$$

where, δ_{ω} denotes the overall weight values added with the input and δ_b indicates the overall bias values added.

The random selection of the weight parameter has an effect on the training process as a whole. So, the parameter needs to be tuned in order for the model to be more accurate. The following words describe the loss or error function that was seen in the RBM:

$$L_s(\omega, a, b) = -\sum_{j=1}^{l} \ln\left[\mathrm{P}\left(v^{(j)}\right)\right] \tag{3.8}$$

where, l indicates the total count of image samples used for training.

3.2.4.2 Weight Update using ROA

ROA finds the best weight parameter for DBN to improve performance. The optimization strategy can determine the neural network weight parameter. ROA in the proposed approach aims to minimize DBN loss function. Writing the fitness function this way:

$$F = \min(L_s) \tag{3.9}$$

The fitness of all solution space search agents is verified. The DBN uses weights throughout the solution space. Weight radius is the most important feature in the proposed weight parameter optimization method. It means that the radius of the weight gets wider over time. A bigger radius means that the weight is larger, and a smaller radius means that the weight is lighter. To find the best weight value out of all the weight values, the limit of the weight parameter is checked. The following conditions are used to figure out the radius of the weight:

a) The search agent changes the position by taking into consideration the weight values of the nearby objects:

$$\Re = \left(\xi_1^n + \xi_2^n\right)^{\frac{1}{n}} \tag{3.10}$$

where, $\xi_1 \; and \; \xi_2$ shows the distances between two weight values, and n shows the search agent variables.

b) The following equation is used to find the lowest weight value based on the training error and the fitness value:

$$\Re = \left(\beta \xi_1^n\right)^{\frac{1}{n}} \tag{3.11}$$

where, β denotes the rate of parameter reduction and controls the balance between exploring and profiting among the search agents. Error is reduced by iterating until the training model weight is optimal.

The features are reconstructed for all the classes separately in the testing process. It is calculated how much the reconstruction failed, and once the errors have been fixed, the voting classifier is used to give each image a vote based on its features. At last, the votes assigned are computed and the features gaining highest votes are labelled into affected class whereas the features gaining lower votes are categorized as unaffected.

3.3 Experimentation and Result Analysis

This section will present the scenario, quantitative measures of performance, and a comparative analysis of the proposed model. The efficiency of the proposed method was confirmed by simulations performed on the MATLAB platform. The algorithm employed specific facial characteristics linked to different emotions to train the classifier and accurately detect facial emotions in individuals. The study utilized the KDEF (Karolinska Directed Emotional Faces) dataset and the Japanese Female Facial Expression (JAFFE) dataset. The KDEF dataset consists of 4900 photos that represent various facial expressions of both male and female people. Among these, 970 images were utilized for training purposes, while 194 images were reserved for testing. The JAFFE dataset is a collection of 213 image samples that were used for emotion recognition. The dataset consists of images of 10 Japanese female participants, and each image has a resolution of 256 × 256 pixels. The dataset categorizes facial expressions as happiness, sadness, fear, surprise, disgust, neutrality, and rage. The analysis used 80% of the photos for training and 20% for testing. Table 3.1 shows formulas for performance metrics, while table 3.2 shows the proposed framework's hyper-parameter settings.

3.3.1 Performance Metrics

The proposed method targets precision, recall, f-measure, accuracy, specificity, and mean square error to identify facial emotion differences and choose the best model.

Table 3. 1: Formulas for performance metrics

SI no.	Performance metrics	Formulas
1.	Precision	$precision = \frac{True_{positive}}{True_{positive} + False_{positive}}$
2.	Recall	$recall = \frac{True_{positive}}{True_{positive} + False_{negative}}$
3.	F1-score	$F_1 score = 2 \times \left(precision^{-1} + recall^{-1} \right)$
4.	Accuracy	$Accuracy = \frac{True_{negative} + True_{positive}}{True_{negative} + True_{positive} + False_{negative} + False_{positive}}$
5.	Specificity	$Specificity = \frac{True_{negative}}{True_{negative} + False\ positive}$
6.	Kappa coefficient	$K = \frac{2 \times (tp \times tn) - (fp \times fn)}{(tp + fp)(tn + fp) + (tp + fn)(tn + fn)}$
7.	Mathew's correlation coefficient (MCC)	$M_C = \frac{(tp \times tn) - (fp \times fn)}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}}$
8.	Mean Square Error	$MSE = \sum_{i=1}^{n} \frac{(\hat{c}_i - c_i)^2}{n}$
9.	False Positive Rate	FPR = FP / (FP+TN)
10.	Recognition Rate	$\mathrm{Re}\, cognition\ rate = \frac{E}{\Phi} \times 100$

Table 3. 2: Proposed hyper-parameters

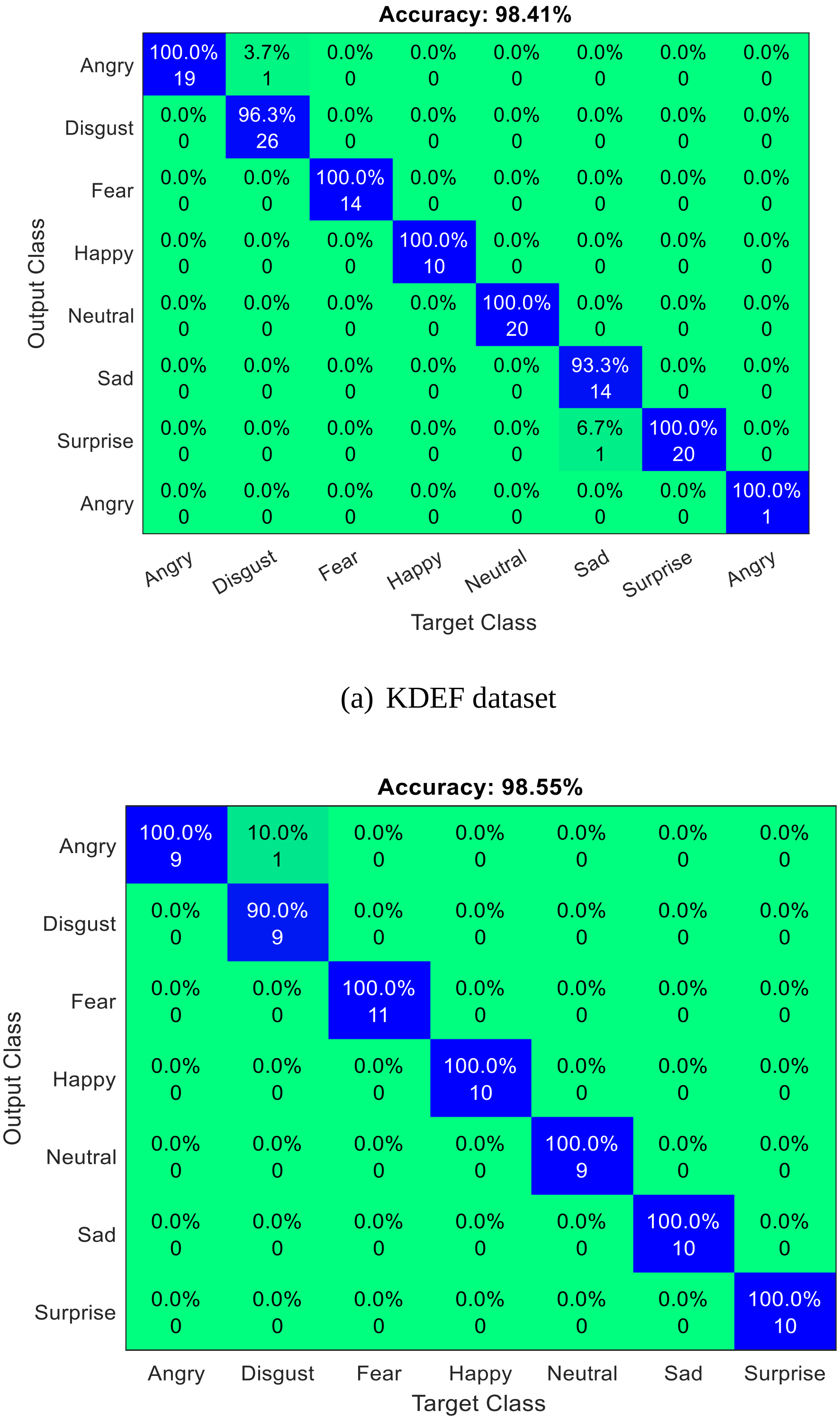
Sl. No	Hyper-parameters	DBRO
1.	Tuning algorithm	ROA
2.	Initial learning rate	0.1%
3.	Max epochs	100
4.	Mini batch size	32
5.	No. of hidden units	10
6.	No. of neurons in the input layer	250
7.	Initial population	100
8.	No. of raindrops	100
9.	Dimension	5
10.	β	10

3.3.2 Performance Evaluation

The proposed DBRO classifier is tested using dataset images and facial feature extraction. The dataset includes seven distinct emotions: happiness, sadness, fear, disgust, anger, surprise, and neutrality. The GF obtains the geometric face points to identify changes in these characteristics, whereas the DBN layers extract variations in characteristics to enhance the learning process. After repeated repetition, The layers take in features or geometric points as input and repeat the same facial features. It tests rebuilt features to see how well the classifier learned from the input characteristics. The DBRO classifier classifies the input photos into seven distinct moods using the provided facial characteristics as input. The framework utilized the analysis of several input facial traits to identify the emotional states of individuals. Geometric feature points defined facial structure appearance features, including skin texture, and intensity values were obtained from these points. The optimization process was used to reduce the dimensions of the feature space, and the classifier training was used to label the features with seven distinct types of emotions. A comparison of proposed model confusion matrices to KDEF and JAFFE datasets shows that the model achieved better precision in accurately classifying emotions based on photos. All 19 furious images in the KDEF dataset were correctly labeled as angry, 26 disgust images were precisely classified as disgust, and 14 sad images were correctly classified as sad. The suggested model achieved an overall classification accuracy of 98.41% for the KDEF dataset.

While looking at different datasets, the proposed algorithm worked better with photos from the JAFFE dataset. The proposed DBRO classification model had a higher AUC than the other models, as shown by a ROC curve analysis of the proposed and current classification methods for the KDEF and JAFFE datasets. The suggested model outperformed existing deep models in terms of AUC values for the JAFFE dataset, with AlexNet achieving similar AUC values to the proposed model.

The study compares the performance of various emotion recognition classifiers using the KDEF and JAFFE datasets. The results unequivocally show that the proposed approach outperforms the other models in terms of precision, recall, f-measure, accuracy, specificity, kappa, false positive rate, Matthew’s correlation coefficient, and error. The proposed model achieved an overall accuracy of 98.41%, with an error rate of 1.59, demonstrating its superiority. The accuracy rating of AlexNet is 97.92%, while ResNet has the lowest accuracy rate at 91.87%. The hybridization of ROA minimizes the impact of mistakes by carefully determining the weight value for the classifier at each iteration. This method reduces the possibility of incorrect classifications. The proposed model achieved the following performance metrics on the KDEF dataset: precision (98.20%), recall (98.78%), f-measure (98.46%), specificity (98.96%), kappa (92.74%), FPR (0.24%), and MC (98.49%).



Accuracy: 98.41%

Output Class \ Target Class	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Angry
Angry	100.0% 19	3.7% 1	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Disgust	0.0% 0	96.3% 26	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Fear	0.0% 0	0.0% 0	100.0% 14	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Happy	0.0% 0	0.0% 0	0.0% 0	100.0% 10	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Neutral	0.0% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 20	0.0% 0	0.0% 0	0.0% 0
Sad	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	93.3% 14	0.0% 0	0.0% 0
Surprise	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	6.7% 1	100.0% 20	0.0% 0
Angry	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 1

(a) KDEF dataset

Accuracy: 98.55%

Output Class \ Target Class	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	100.0% 9	10.0% 1	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Disgust	0.0% 0	90.0% 9	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Fear	0.0% 0	0.0% 0	100.0% 11	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Happy	0.0% 0	0.0% 0	0.0% 0	100.0% 10	0.0% 0	0.0% 0	0.0% 0
Neutral	0.0% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 9	0.0% 0	0.0% 0
Sad	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 10	0.0% 0
Surprise	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 10

(b) JAFFE dataset

Figure 3. 5: The confusion matrix of the suggested technique is presented for two datasets

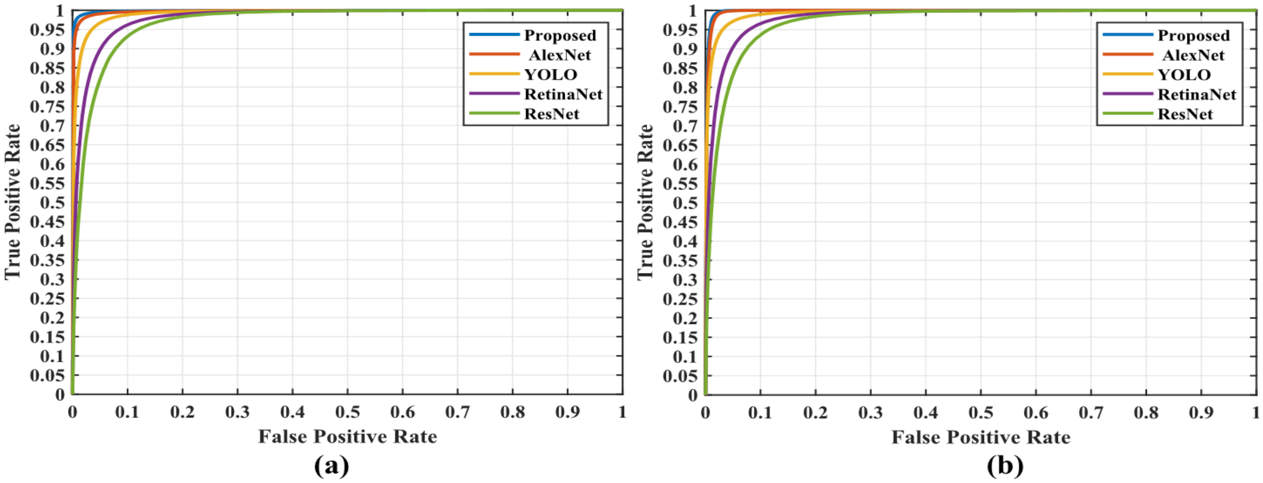


Figure 3.6: ROC curves are shown for both the proposed and existing classification methods (a) the KDEF dataset and (b) the JAFFE dataset.

The hybridization of ROA reduces the influence of error by precisely choosing the weight assigned to the classifier for each event. The suggested approach for the JAFFE dataset achieves an overall accuracy of 98.55% with an error value of 1.45. AlexNet gets a higher accuracy rate of 98.11%, while ResNet has a comparatively lower accuracy rate of 92.06%. The DBRO model, when applied with the ROA algorithm, achieves higher precision with zero misclassifications. This graph presents a comparison of the precise emotion recognition capabilities of the proposed and existing models for both datasets. The ideal characteristics of the model and its classification based on context offer a greater level of context for understanding emotions. The JAFFE dataset shows greater values compared to the KDEF dataset. The hybridization of ROA improves model training, allowing it to learn features and classify without errors. The indicated model, AlexNet, outperformed existing models by achieving better accuracy values. It attained an overall accuracy of 98.41% for the KDEF dataset and 98.55% for the JAFFE dataset. The feature selection process produced the best classifier training characteristics, eliminating dimensionality issues. The model's objective-based fitness function found optimal features. In each iteration, MOSOA used training and testing samples to assess fitness, observing small gains in fitness ranging from 0 to 40. With each successive iteration, the model improved faster convergence and yielded ideal results. This strategy reduces the probability of incorrect classifications. The proposed model achieved the following performance metrics on the KDEF dataset: accuracy of 98.20%, recall of 98.78%, f-measure of 98.46%, specificity of 98.96%, kappa of 92.74%, false positive rate (FPR) of 0.24%, and Matthew’s correlation coefficient (MCC) of 98.49%.

3.3.3 Ablation Study

Multiple ablation experiments were performed to evaluate each part of the proposed method. The models were tested in four modules: Module 1, Module 2, Module 3, and Module 4. Each module's accuracy is evaluated separately to determine the significance of the proposed work's four components. All four phases of the proposed work are defined in Module 1. In Module 2, all four phases of the suggested framework are shown. Still, the classification is done with no modification to any of the parameters. Module 3 includes pre-processing, feature extraction, and classification. Module 4 extracts and classifies features.

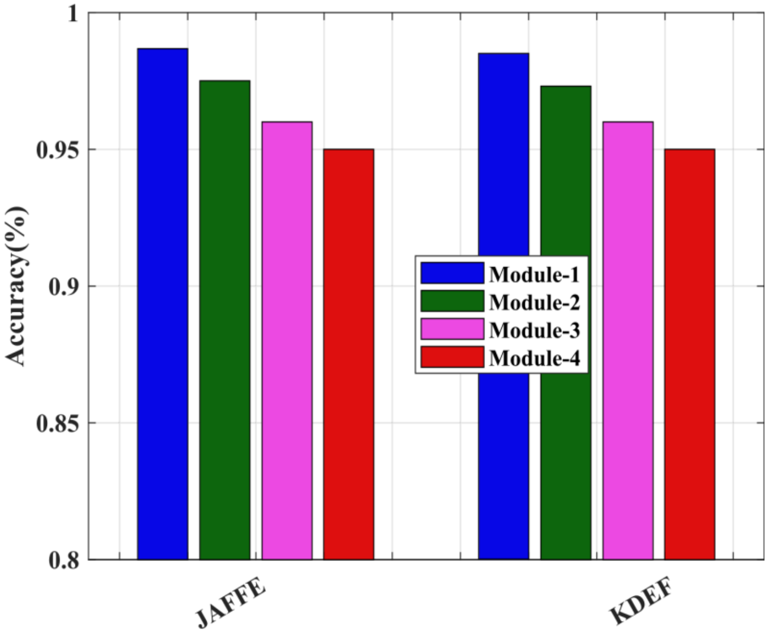


Figure 3. 7: Ablation study for the proposed approach.

The complete performance of each module in the model that is suggested is illustrated in Figure 3.7. The graph shows that each step of the suggested method makes an equal contribution to classifying emotions. Module 4 performs worst because it only consists of feature extraction and classification. This model extracts image features without processing. Because of this, the images' distortions made the classification less accurate.

In the same way, Module 3 skips the feature selection phase, which makes the features more complex. It took longer to train this model, and it had a low accuracy rate. Module 2 gave the best results compared to Modules 3 and 4, but this model changed the error rates, which made the accuracy rate lower. Module A, which includes all four steps of the proposed approach, gives the best performance.

This study explains why each part of the proposed model is important for recognizing emotions. Images that are noisy can't be used directly for classifying and extracting features. The features that are extracted from these noisy images have extra information that makes classification less effective and efficient. The long training time is caused by the large number of features that are extracted from each image, which makes time complexity problems. Module 4 shows how important the pre-processing and feature selection stages are because of this. The MOSOA algorithm quickly found the best features in the feature space, which helped with problems caused by the large number of dimensions. Why is the feature selection phase important in the proposed model? The difference in performance between Modules 2 and 3 shows this. There are also small differences in how well Module 1 and Module 2 work. Misclassifications happen because of the effect of error rates in Module 2. Performance went down. The ROA algorithm helped get the best results by figuring out the right weights for the classifier during the training phase. The algorithm always shows the best value for learning, no matter how many times it is run. Because of this, emotion recognition is now more accurate across both datasets. The ablation tests demonstrate that all four steps of the suggested method are equally important for detecting emotions.

3.3.2 Analysis of inference time:

The inference time is examined to assess the proposed model's emotion classification performance and efficiency. The MOSOA algorithm reduces model inference time by providing only the best classifier training features. Due to improved training, the proposed model produced inference results quickly. The proposed and existing models' inference times are examined below:

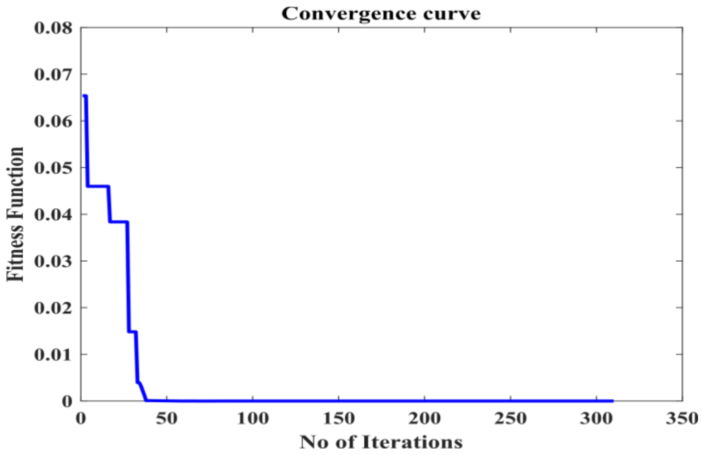


Figure 3. 8: The convergence curve depicts the progress of the MOSOA algorithm for feature selection.

Figure 3.8 depicts the convergence of the proposed MOSOA feature selection algorithm. On the x-axis, multiple iterations are plotted against the fitness function on the y-axis. The graph shows faster convergence of the proposed model with more iterations. The feature selection phase selected optimal classifier training features without dimensionality issues. This algorithm gave the model optimal features that accurately defined emotions. The model's fitness function identified objective-optimal features. MOSOA assessed each training and testing sample iteration's efficacy. During iterations ranging from 0 to 40, a marginal enhancement in fitness is observed; however, as iterations increase, the MOSOA converges more rapidly and yields optimal outcomes.

Table 3. 3: Analysis of interpretation time for the proposed and current classifiers for a single image from the KDEF and JAFFE datasets

Methods	KDEF (sec)	JAFFE (sec)
ELM	1.26	1.98
VGG-16	0.59	0.46
AlexNet	0.25	0.22
ResNet	0.11	0.09
DFLNet	0.09	0.06
Proposed	**0.08**	**0.05**

Table 3.3 shows the inference time for classifying a single image using the proposed and existing classifiers on KDEF and JAFFE datasets. The proposed model takes 0.08 seconds to identify a KDEF image and 0.05 seconds for a JAFFE image. Superior results and reduced time complexity were achieved by the proposed model. ResNet took 0.25 seconds for the KDEF dataset and 0.22 seconds for the JAFFE dataset to infer a single image. AlexNet reached optimal performance with an average inference time of 0.11 seconds for a KDEF image and 0.09 seconds for a JAFFE image. The proposed model processed images quickly using feature selection and parameter tuning. Thus, the proposed model is suitable for emotion classification to improve accuracy.

The proposed approach for emotion recognition has four distinct modules: Module 1, Module 2, Module 3, and Module 4. The modules are divided into four distinct phases, with each step providing a similar contribution to emotion classification. Module 4 offers the lowest level of performance as it primarily focuses on feature extraction and classification stages, leading to visual distortions and diminished accuracy. Module 3 excludes the process of feature selection, which results in a higher number of features and longer training time.

Module 2 produces ideal outcomes but also affects error rates, which decreases accuracy. Module A offers the most optimal performance as it combines all four phases. The analysis emphasizes the significance of each phase in the emotion recognition process. Images with excessive noise are unsuitable for feature extraction and classification, as they introduce unnecessary information and prolong training duration. Module 4 emphasizes pre-processing and feature selection. The MOSOA algorithm effectively identifies the most suitable characteristics from the feature space, thus minimizing issues related to dimensionality. Modules 2 and 3 demonstrate differences in performance, highlighting the importance of selecting the appropriate features. Module 1 and 2 show slight discrepancies because of the error rates in Module 2, therefore resulting in incorrect classifications. The ROA method achieves optimal results by selecting classifier weights during training. The ablation studies indicate that all four phases of the proposed technique contribute equally to emotion perceptions.

Table 3. 4: Performance values of the classification models for the JAFFE dataset

Methods	Precision	Recall	F-measure	Accuracy	Specificity	Kappa	FPR	MC	MSE
	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
ELM	87.89	87.12	88.35	88.21	90.50	80.26	8.97	81.46	8.69
VGG-16	89.56	87.64	89.74	90.10	92.56	81.70	7.02	84.23	7.98
AlexNet	91.64	89.79	90.70	92.06	93.08	83.88	6.12	83.90	7.94
ResNet	96.88	98.24	97.55	98.11	97.21	96.15	1.99	96.15	1.89
DFLNet	97.21	98.27	97.89	90.97	97.84	96.52	1.65	96.91	1.73
DBRO (For JAFFE)	98.07	98.57	98.25	98.55	98.96	94.08	0.24	98.31	1.45
DBRO (For KDEF)	98.20	98.78	98.46	98.41	98.96	92.74	0.24	98.49	1.59

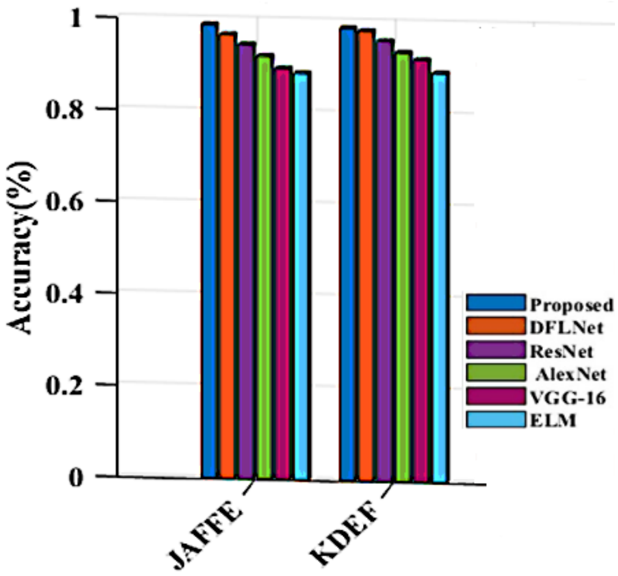


Figure 3. 9: Accuracy comparison of the proposed and existing models.

Figure 3.9 compares the proposed and current emotion recognition models for the KDEF and JAFFE datasets. A graph shows that the proposed model outperformed the others. The best features have been chosen for classification, which improves the model's capacity to distinguish between different emotions. By improving model training, the hybridization of ROA allows the model to classify features accurately and with low error rates. DFLNet achieved greater accuracy than the other models that were assessed. The overall accuracy of the suggested model is 98.55% for the JAFFE dataset and 98.41% for the KDEF dataset. ELM has the worst performance and most misclassifications.

3.3.4 Analysis of Implication Time

The proposed emotion classification model is evaluated for performance and efficiency. The MOSOA algorithm plays a significant role in reducing inference time, as it provides optimal features for training. The model's inference time has been significantly reduced due to improvements. Table 3.5 shows the proposed and current classifiers' KDEF and JAFFE inference times. For the KDEF dataset, the suggested model classifies images in 0.08 seconds and for the JAFFE dataset, 0.05 seconds. The proposed method beats previous models by delivering optimal results while minimizing time complexity. ResNet has a longer inference time compared to AlexNet. On average, AlexNet processes a KDEF image in 0.11 seconds and a JAFFE image in 0.09 seconds, yielding optimal results. The proposed framework is appropriate for emotion classification, as it offers a greater level of accuracy in classifying emotions.

Table 3. 5: An examination of inferences using the proposed and current classifiers for a single image in the KDEF and JAFFE datasets.

Methods	**KDEF**	**JAFFE**
ResNet	**0.25**	**0.22**
RetinaNet	**0.19**	**0.17**
YOLO	**0.14**	**0.13**
AlexNet	**0.11**	**0.09**
Proposed	**0.08**	**0.05**

3.4 Discussion

The DBRO-based framework for recognizing emotions on faces did better in all performance criteria that were looked at. An evaluation of the many stages of the model has shown that each stage is essential for enhancing performance. The Joint Bilateral Filter (JBF) improves image quality and retains edge information during pre-processing. Smoothing features during feature extraction improves efficiency. An analysis of the HOG and GF features obtained during feature extraction provides further insights into individuals' emotions. The feature selection phase addresses classification dimensionality issues to improve accuracy. The hybridization procedure increases training phase weight values using ROA-optimal values. The proposed MOSOA-based feature selection method reduces the training phase feature count, leading to enhanced training accuracy and reduced inference time during testing. The model consistently showed neither over-fitting nor under-fitting problems when tested on both datasets. It suggests the model is well-suited for accurately classifying facial emotions, even when given new, unseen examples. The usual DBN model exhibits deficiencies in terms of its robustness in classification and its tendency to misclassify images with low accuracy. The suggested model improves the conventional DBN by using improved weight values and efficiently improving the classification process. The DBRO approach efficiently identifies face landmarks, resulting in improved training efficiency using stacked Limited Boltzmann Machines (RBMs). The model achieves convergence in a small number of iterations, leading to a low level of computational complexity. The suggested approach demonstrates greater precision for recognizing emotions compared to current state-of-the-art techniques. The DBRO model provides optimal outcomes by using feature selection and weight optimization techniques, unlike other models that require larger datasets and involve more time complexities. The training model adopted in [84] improves classification accuracy, but it yields unsatisfactory outcomes when the number of training samples is increased. In summary, the proposed model demonstrates its efficacy and effectiveness in identifying emotions. It also requires low training time and can learn features from newer data with good generalization.

3.5 Summary

This chapter introduces a new method for automatically classifying facial emotions using a hybrid Deep Belief Rain Optimization (DBRO) approach, inspired from Deep Belief Network and Rain Optimization Algorithm. The model enhances image quality through pre-processing, identifies significant features such as geometric and appearance-based ones using HOG and GF, decreases dimensionality through choosing features with MOSOA, and finally classifies the selected features using an optimum weight parameter. The experimental findings show that the technique effectively classifies photos into seven distinct moods using input information, resulting in a 97% level of accuracy. We envisioned making better models that make training more stable overall, especially when trained for long periods using different sets of data.

Chapter 4: Bi-directional Elman Neural Network for Facial Emotion Recognition

In the previous chapter, we foresaw better models that would add training stability overall, especially with different ways of training for more extended periods using different sets of data. Besides this, we set ourselves the goal of developing more fine-tuned models to be able to enhance the general stability of training, even if training for greater lengths of iterations with different samples of data. The traditional ENN's primary problem is limited classification context. This chapter introduces a novel Bi-ENN (Bi-directional Elman Neural Network) approach to achieve optimal facial emotion classification results by overcoming established neural network challenges like unavailable context during training. The Bi-ENN can be used in a FER system to classify individual emotions, select pertinent features, and mitigate the dimensionality issue in a stable environment.

4.1 Introduction

Neural network structure plays an important role in discerning variations in facial characteristics for facial emotion recognition. Feature extraction is an essential classifier with suitable feature vectors for effective categorizations. Most current FER models rely on deep learning, as these methods have proven to be more efficient in classification compared to previous techniques. In this chapter, the proposed technique seeks to develop an optimized face Expression Recognition (FER) system that effectively distinguishes and categorizes face features. Current face emotion classification models predominantly rely on deep learning techniques. However, these models encounter difficulties in effectively distinguishing input information, leading to decreased accuracy rates. To address this issue, we propose a novel neural network called the bidirectional Elman neural network (Bi-ENN) as an alternative to current deep learning models. Using the bidirectional long short-term memory (Bi-LSTM) neural network, the Bi-ENN enhances the training context by incorporating both forward and backward training. The objective of the proposed approach, Enhanced Battle Royale Optimization (EBRO), is to refine the feature selection process by minimizing the dimensionality and time complexity of the input. The primary application of the Bi-ENN is the categorization of individuals' emotions. This novel methodology seeks to enhance the precision and effectiveness of facial emotion classification [85].

In this chapter, A novel Bi-ENN (Bi-directional Elman Neural Network) is introduced to achieve optimal classification results in the facial emotion domain. The Bi-ENN is used in a FER system to classify individuals' emotions, select pertinent features, and minimize the dimensionality issue in a stable environment. Simulations prove that the proposed neural network outperforms other classification algorithms in accuracy. The model effectively discerns emotions because of the enhanced training context. In addition, the proposed model has been verified using two distinct datasets, yielding nearly identical outcomes. The model stays invariant to the distribution or magnitude of the data. The implementation achieved an accuracy of 98.57% on the JAFFE dataset and 98.75% on the CK+ dataset.

4.2 Proposed Method

A novel FER system for emotion classification based on deep learning models is proposed. The system offers higher accuracy results and is designed for optimal performance. The four major phases of the architecture are as follows:

a. Pre-processing
b. Feature extraction
c. Feature selection
d. Classification

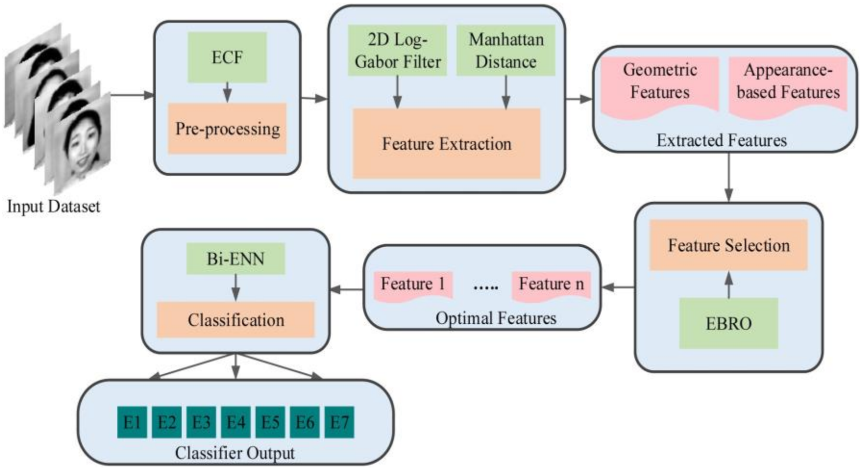


Figure 4.1: The proposed framework.

After extracting geometric facial points and appearance-based features, the proposed framework pre-processes facial images to remove noise and distortions. The classification phase classifies images using the proposed NN after selecting relevant features and reducing their dimensionality. The framework's architecture is illustrated in Figure 4.1. As we can see from the above figure, the implied strategy aims to evaluate photos of the facial emotions dataset by removing any interference or alterations using the ECF technique. The system extracts prominent geometric and appearance-based characteristics from each input image, using face cues that represent various emotions while considering changes in skin and texture. The EBRO method is employed to extract the most beneficial and distinctive characteristics, using chaos theory to find highly discriminative features. The method analyzes forward and backward ENN model training for accurate prediction results. The proposed FER system undergoes extensive testing to demonstrate its effectiveness and efficiency using well-known FER datasets such as JAFFE and CK+.

4.3 Pre-processing

The process of facial expression recognition involves several crucial steps, including data acquisition, preprocessing, and deep learning model training. Preprocessing approaches are crucial for improving the accuracy of face expression recognition models. Gaussian smoothing is a crucial preprocessing technique that improves image quality by reducing noise. It achieves this by averaging neighboring pixels, resulting in a smoother image. Median filtering is another effective method, replacing each pixel's value with its neighborhood's median to reduce noise and preserve edges. Wavelet denoising breaks the image into frequency bands to remove noise. Histogram equalization and contrast stretching improve image contrast and quality, while normalization transforms the image's intensity values to a specific range. Interpolation methods, such as nearest neighbor, bilinear, and bicubic, are employed to resize the image while maintaining image quality. The gray world assumption and white balance techniques are used to enhance color balance and improve overall image quality.

In this research, the preprocessing stage includes three steps: resizing the image to a uniform size, converting the image to grayscale to reduce color information and enhance contrast, and normalizing the image to ensure all pixel values are within a specific range, further enhancing contrast and improving image quality. By applying these preprocessing techniques, the images can be improved in terms of noise reduction, contrast enhancement, and edge detection, ultimately leading to better performance in facial expression recognition systems.

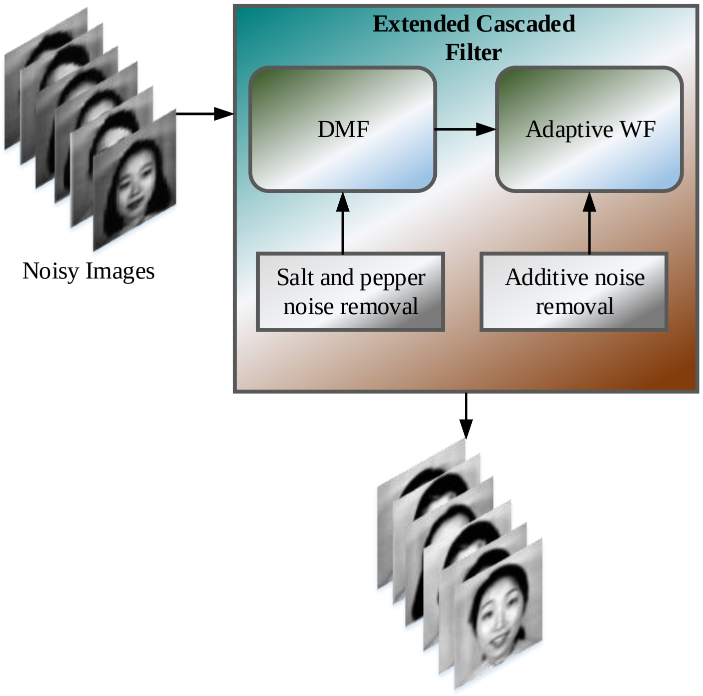


Figure 4. 2: Block diagram of ECF for pre-processing.

First, the DMF removes salt and pepper noise from images. DMF filters examine images using sliding windows to locate salt and pepper noise. The median value of pixel noise is used. This is followed by the adaptive Wiener filter to reduce blurring and retain edge information. ECF de-noises input images using the following methods:

Step 1: A 2D sliding window is initially selected to acquire the pixel values of the image.

Step 2: Among the pixel values in the image, one pixel value ($\wp_{xy}$) is chosen and the value is compared between the pixel limits to identify the noisy pixels (i.e. if $\wp_{xy}$ <0 or if $\wp_{xy}$ >255, then the pixel value is considered as noisy pixel).

Step 3: If a pixel value conflicts with both conditions, it is classified as noiseless and remains unprocessed.

Step 4: The median value of the sliding window, excluding 0 and 255, is calculated for the noisy pixel values, and this value is utilized to replace the affected pixel values.

Step 5: The sliding window is subsequently shifted to the next pixels in the image.

Step 6: Steps 2 to 5 are repeated until all pixels of the image have been processed.

Step 7: After the removal of salt and pepper noise, the images are processed through the adaptive Wiener filter to mitigate blurring and retain edge information. The expression for the Wiener filter can be expressed as follows:

$$I(p,q) = \mu + \frac{\sigma_\ell^2 - \sigma_g^2}{\sigma_\ell^2}\left(n(p,q) - \mu\right) \tag{4.1}$$

where, p,q indicates the pixel values of the image, $n(p,q)$ is the original noisy image, $I(p,q)$ is the de-noised image, μ is the mean of each pixel, σ_ℓ is the local variance and σ_g is the global variance.

4.4 Feature Extraction

The second step of the suggested design includes the feature extraction phase, which involves extracting face feature points based on geometry and appearance characteristics. The identification of geometric features is initially achieved through the application of the Manhattan distance metric, while the extraction of appearance-based features is performed using the 2D log-Gabor filter.

4.4.1 Geometric Feature Extraction

This technique uses the Manhattan distance measure to extract geometric characteristics from facial photos, specifically focusing on the eyes, nose, and mouth. This strong distance metric is more dependable than Euclidean distance metrics and is not affected by outliers. It aids in collecting insignificant details from even more minute facial features. The extraction technique involves following precise steps to extract the face features.

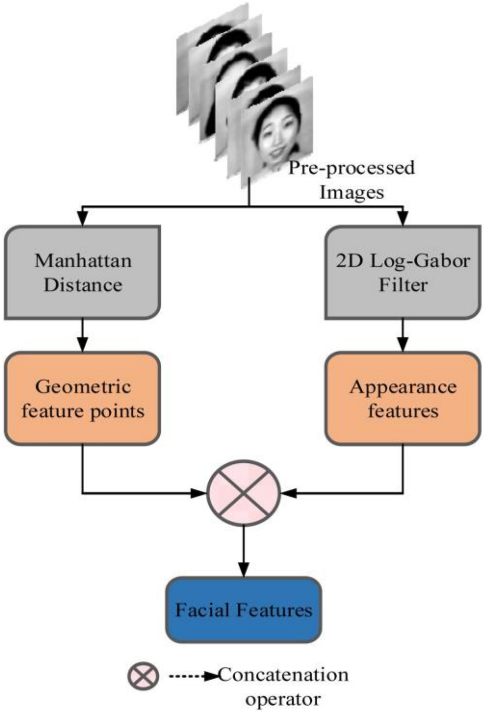


Figure 4.3: Block diagram of the proposed feature extraction phase.

Facial variations may cause geometric feature points to ignore skin changes and textures in photos, which can impact feature extraction and result in inaccurate classification outcomes. Thus, appearance-based features are retrieved to collect both prominent and subtle characteristics.

4.4.2 Appearance-based Feature Extraction

The proposed method uses appearance-based feature extraction to extract texture information from photos, allowing efficient classification. The 2D log-Gabor filter, an enhanced iteration of the conventional Gabor filter, facilitates this procedure. This filter is highly versatile and proficient at identifying intensity values and edge information, as well as sensitive to frequencies, scaling, and orientation. It is used to extract major and minor features from images, enhancing classification efficiency.

4.5 Feature Selection

The proposed work includes a step of picking features, where significant facial expression elements are picked to decrease the dimensionality of the input to the classifier and the complexity of the processing. The approach employs an integrated battle royale optimization algorithm to identify the most beneficial characteristics from the available feature space. But before that, we need to know about the classic Battle Royal Optimization approach.

4.5.1 Description of Classic BRO

Battle Royale Optimizations (BRO) is a meta-heuristic algorithm influenced by the Battle Royale video game. The approach is based on population, considering players who possess similar levels of skill and resources. Players are confined to a secure zone, while those who are less powerful are considered weak. The ideal solution is chosen based on the player's ability to survive, as the game only allows for one player to remain alive. The solution space's best feature vector can be selected using this method [86].

4.5.2 Integrated Battle Royal Optimization (IBRO) for Feature Selection

The algorithm finds the most advantageous characteristics from the set of potential solutions, with a focus on the individuals inside the search space who are affected by harm. The optimal characteristics of the winning players offer additional insight into facial expressions, guaranteeing a very efficient search. Integrating chaos theory into the BRO algorithm can enhance its performance by accurately representing population variety, which is highly responsive to beginning conditions. Introducing random behavior features can enhance the diversity of a population. It improves the capacity to recognize the globally ideal solution inside the problem domain. The traditional BRO algorithm employs a sinusoidal map to represent the exploitation equation, wherein players go to random locations to engage in combat with adversaries. It enables the relocation of characteristics with reduced data on facial expressions to acquire globally optimal features. The enhanced application helps in the identification of globally optimal features, improving the recognition of relevant elements while ignoring less informative ones [87].

The objective function of the proposed feature selection phase can be described as follows:

$$Min\ F = \Psi + \frac{(extracted\ features)}{|\Phi|} \tag{4.2}$$

where, Ψ denotes the classification error and Φ denotes total features found in the whole dataset.

The improved iteration of the BRO algorithm can be formulated using chaos theory to simulate population diversity. This theory depends heavily on initial conditions and how random behavior in the BRO algorithm can increase population diversity. The global optimal solution in the problem space is identified by this enhancement. Chaos theory's discrete dynamic function is:

$$\xi M_{i+1}^{j} = f\left(\xi M_{i}^{j}\right); \quad j = 1, 2, ...\lambda \tag{4.3}$$

where, λ indicates the map dimension and $f\left(\xi M_{i}^{j}\right)$ is the model generator function based on different maps.

Traditional BRO algorithms use sinusoidal maps to model exploitation equations. Broken players in the search space fight in a random area. Thus, the search space's least informative facial expression features are in a sinusoidal modified to produce the most effective features. The sinusoidal chaotic map is:

$$\begin{aligned} &m_{\lambda+1} = \alpha m_{\lambda}^{2} \sin\left(\pi m_{\lambda}\right); \\ &m_{0} \in [0,1]; \ \alpha \in [0,4] \end{aligned} \tag{4.4}$$

where, α is the control parameter.

Improvements in exploitation help find globally optimal solution space features. All facial expression-related features are easily identifiable, while least-relevant ones are removed. The EBRO algorithm flowchart is shown in Figure 4.3.

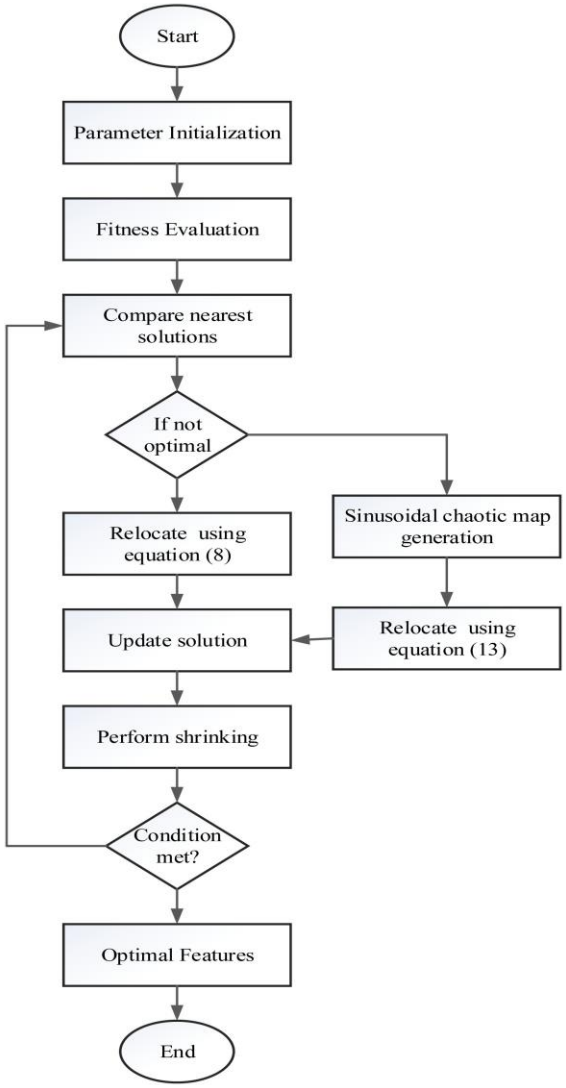


Figure 4. 4: Proposed Enhance Battle Royal Optimization (EBRO) algorithm flowchart.

4.6 Classification

The suggested methodology starts with the neural network receiving features from the feature selection step. A new neural network improves global stability and classification performance. The main issue with conventional ENNs is inadequate classification context. Figure 4.4 illustrates the configuration of the traditional ENN.

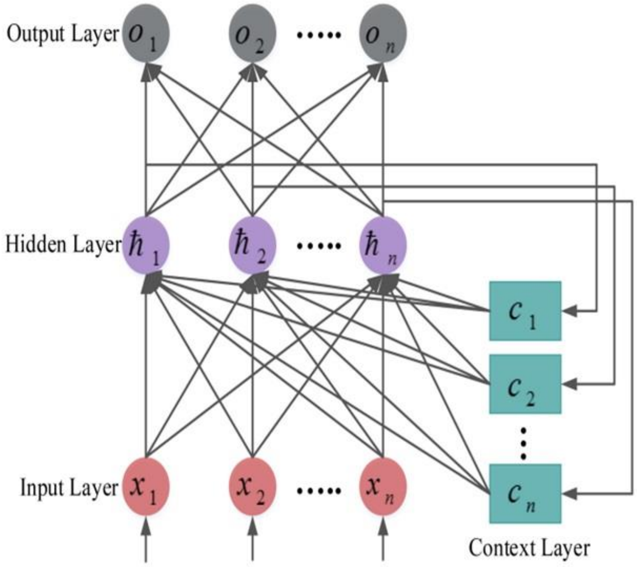


Figure 4. 5: Conventional Elman Neural Network.

The method suggested is to use features obtained during the feature selection phase as input for a unique deep Neural Network (NN). The Elman neural network (ENN) is suggested to achieve more accurate classification results efficiently. The conventional ENN has constraints in terms of the contextual information it may use for classifications. In contrast, the proposed Bi-ENN aims to address this issue by training two ENNs in both the forward and backward directions. The Bi-ENN algorithm effectively retains historical and prospective data, resulting in more accurate and precise emotion prediction [88].

A bi-directional neural network is proposed to identify expressions using the selected face features as input, both in the positive and negative perspectives. It improves the context layer of the ENN by allowing it to retain knowledge about previous and upcoming input features. The input layer sends the feature vector to the hidden layer, which combines weight and bias. The buried layer captures the most relevant characteristics appropriate for classification. The suggested neural network is discriminative, accurately identifying the correct class labels for the given feature.

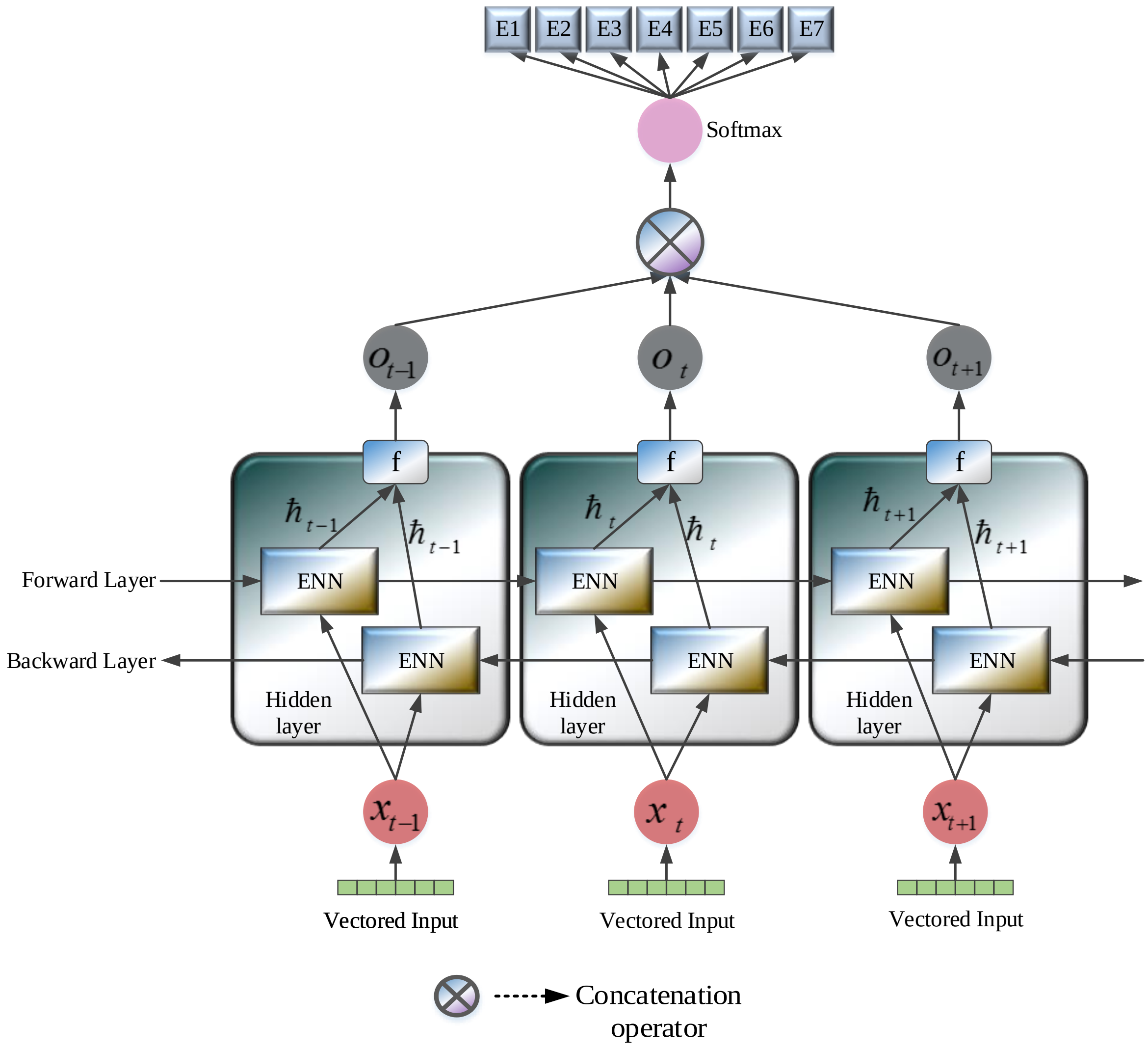


Figure 4. 6: Bidirectional Elman Neural Network (Bi-ENN).

A novel bi-directional Elman neural network (Bi-ENN) is proposed to address the limitation of context unavailability and enhance classification performance. The Bi-ENN is derived from the bi-directional long short-term memory (Bi-LSTM) neural network. The conventional Bi-LSTM is subject to overfitting and exhibits stability during extended iterations. Consequently, the Bi-ENN is introduced, wherein this neural network trains two ENNs in both forward and backward orientations to enhance the context that is available for training. This implementation improves the context layer in the conventional ENN by retaining both historical and future information. This results in an outstanding result characterized by high accuracy and global stability. The design of the proposed Bi-ENN is illustrated in Figure 4.5.

The proposed neural network for expression recognition in forward and backward orientations uses the selected facial features. This method improves ENN's context layer by preserving historical and prospective input feature information. The neural network's input layer sends the feature vector to the hidden layer, which adds weight and bias. The hidden layer finds the best classification features. The proposed neural network discriminates features to classify them correctly.

The implementation of the concealed layer can be described as follows:

$$\hbar_t = f_\hbar\left(\omega_\hbar x_t + \vartheta_\hbar \hbar_{t-1} + b_\hbar\right) \tag{4.5}$$

where, $\hbar_t$ indicates the hidden layer output, $f_\hbar$ is the activation function, x_t indicates the input data, $\omega_\hbar$ *and* $\vartheta_\hbar$ are the weight matrices added to the input and $b_\hbar$ indicates the bias vector. The ENN includes a context layer that uses self-referential feedback to preserve information from prior activations. The context layer can be established based on the subsequent formulation:

$$c_{t-1}^{k} = \hbar_t^{j} \tag{4.6}$$

where, c_{t-1}^{k} indicates the output of k^{th} context layer and $\hbar_t^{j}$ indicates the j^{th} hidden layer. The overall output of Bi-ENN can be given as follows:

$$o_t = f_o\left(\omega_o\left[\vec{\hbar}_t, \overleftarrow{\hbar}_t\right] + b_o\right) \tag{4.7}$$

where, $\vec{\hbar}_t$ *and* $\overleftarrow{\hbar}_t$ are the hidden layer output in forward and backward directions, o_t indicates the output of ENN, ω_o indicates the weight matrix, f_o indicates the activation function and b_o indicates the bias vector. ReLU is more efficient than other activation functions, consequently the suggested neural network uses it. At the neural network's end, the SoftMax function classifies the feature vectors into seven emotions.

The loss function of the proposed neural network is defined by the following mean square error (MSE) formula:

$$L_f = \frac{1}{N}\sum_{i=1}^{N}\left(o_i^{ex} - o_i^{ac}\right)^2 \tag{4.8}$$

where, N indicates the number of samples chosen for training, o_i^{ex} indicates the expected classification output and o_i^{ac} indicates the actual classification output of the NN.

4.7 Experimentation and Result Analysis

The NN and FER systems were tested using JAFFE and expanded CK+ images. Bi-LSTM and ENN neural networks evaluate Bi-ENN recognition. Assessment of all FER system effects. We use JAFFE and CK+ to train models. ECF uses Manhattan distance and 2D log-Gabor filter to extract geometric and appearance-based photo characteristics. From extracted data, 29 features were selected for categorization. The suggested Bi-ENN model takes all selected features for classification, transforms them into vectors, and uses them as input. The neural network is trained using input characteristics, and the neurons in the hidden layer extract distinctive properties to assign labels to feature vectors belonging to seven different emotion classes. Forward and backward context layers preserve historical and future data to provide accurate projections of class labels. Evaluations of the suggested framework's performance are provided.

4.7.1 Simulation

The FER system and Bi-ENN have been evaluated using MATLAB in several experimental settings. The JAFFE and CK+ databases are utilized for assessments, encompassing diverse facial expressions of individuals with suitable class designations. The JAFFE dataset comprises 213 photos of 10 Japanese girls, each with a resolution of 256 × 256 pixels, depicting seven distinct facial moods. The CK+ dataset comprises eight distinct facial emotions, with a total of 123 participants and 593 photos, all captured at a resolution of 640 × 490 pixels. The suggested model is initially assessed using these datasets. The data is divided into an 80-20 ratio, with 80% allocated for training and 20% allocated for testing.

4.7.2 Performance Metrics

The tuning algorithm selected for this model is Adam, a popular optimization algorithm known for its efficiency and adaptive learning rate. The initial learning rate is set at a very low value of 0.01%, which helps in fine-tuning the learning process to prevent overshooting the minimum during optimization. For the training process, a mini-batch size of 27 is used, meaning that the model processes 27 samples at a time before updating the weights. The training is conducted over a maximum of 500 epochs, allowing the model sufficient iterations to learn from the data. The input to the model consists of 213 units, and the input size is specified as 29, likely referring to the dimensionality or the number of features in each input sample.

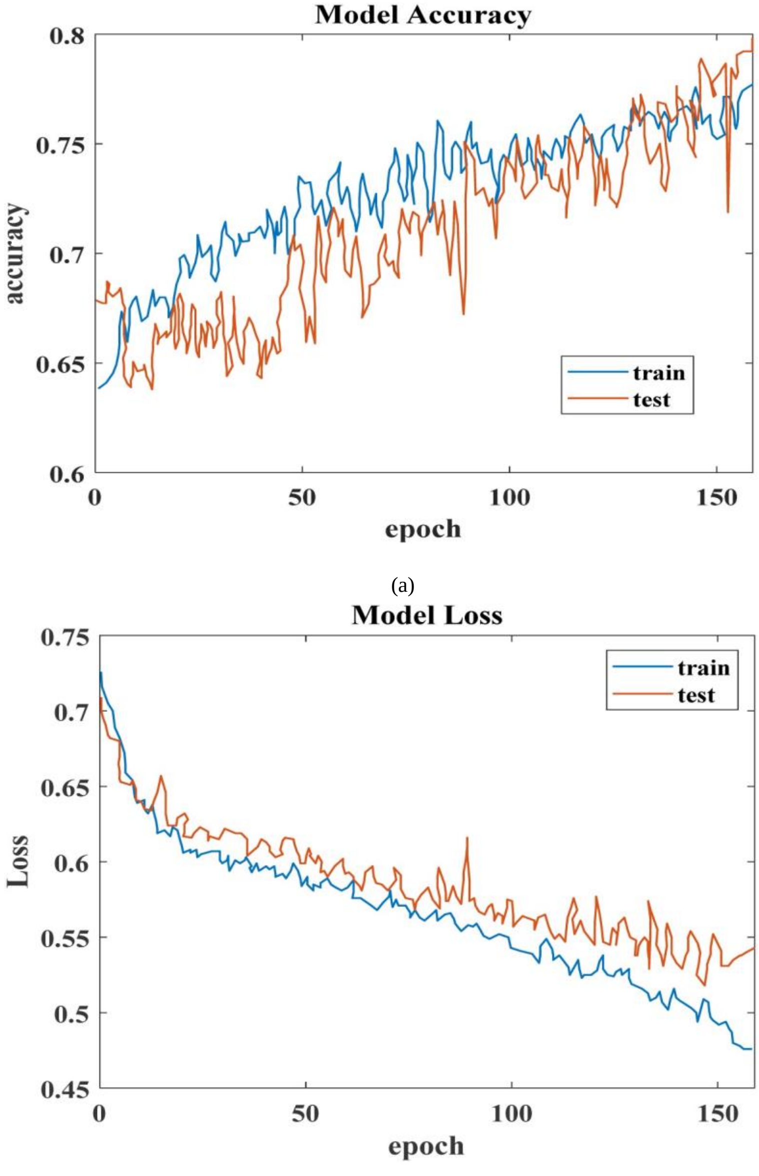
Within the network, there is a hidden layer composed of 100 neurons. This hidden layer configuration helps the model capture complex patterns in the data. The hidden neuron size, also set to 100, indicates the size or number of parameters associated with each hidden neuron. Finally, the model has a single output layer, which is common in many classification and regression tasks where the model needs to output a single prediction value or class label. These hyper-parameters are crucial as they determine the structure and the learning behavior of the Bi-ENN, significantly impacting its performance and accuracy in tasks such as Feature Extraction and Recognition (FER). The system utilized for running the proposed Bi-ENN framework is equipped with an Intel Core i3-2120 CPU, operating at a clock speed of 3.30 GHz. This processor belongs to Intel's third generation of Core processors, providing a balance of performance and energy efficiency suitable for various computational tasks. The system has 8.00 GB of installed RAM, which is adequate for handling the data and processing requirements of the Bi-ENN model during the training and evaluation phases.

4.7.3 Performance Evaluation

This section presents empirical evaluations of the neural network (NN) and facial expression recognition (FER) system under consideration using various sophisticated methodologies and architectures. The algorithms used in model analysis include bee colony optimization, particle swarm optimization, whale optimization method, cat swarm optimization, ICSO, neural networks (NN), support vector machines (SVM), ensemble-based neural networks, and ensemble-based support vector machines. The present algorithms' values are derived from the ICSO24 model. An analysis is conducted on the interpretation of the systems on both datasets.

4.7.4 Analysis of Bi-ENN for FER

The Bi-ENN is tested on two datasets to analyze differences in performance in input samples. The results demonstrate that the suggested Bi-ENN yields almost identical outcomes for both datasets, as the context for the neural network has been enhanced. It aids the neural network in understanding the distinction among input characteristics.

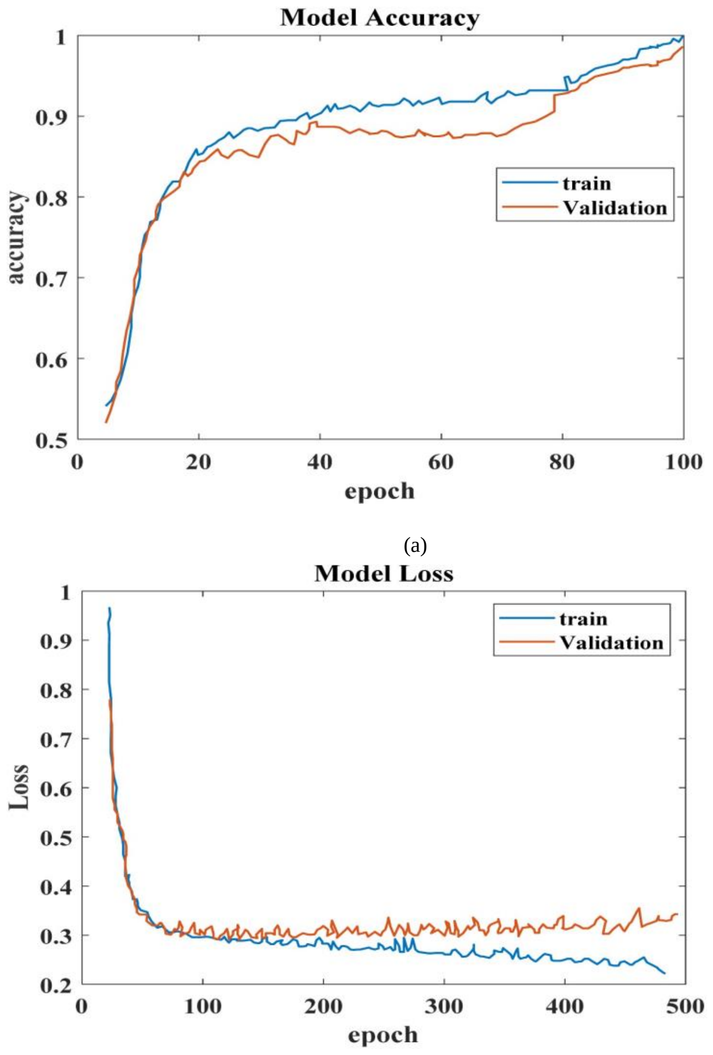


(a)

(b)

Figure 4. 7: (a) Bi-ENN's accuracy graph for the train and test samples on JAFFE and (b) a loss graph for the train and test samples on JAFFE.

The Bi-ENN algorithm uses photos from the JAFFE dataset as input and accurately classifies the feature vectors using a neural network. The visual representations of the Bi-ENN demonstrate that the accuracy comparison between the training and testing samples is comparable. As the number of epochs increases, the performance of the model significantly improves. Following 150 epochs, the neural network attained an accuracy of 80%. The Bi-ENN attained an overall accuracy rating of 98.57% on the JAFFE dataset.

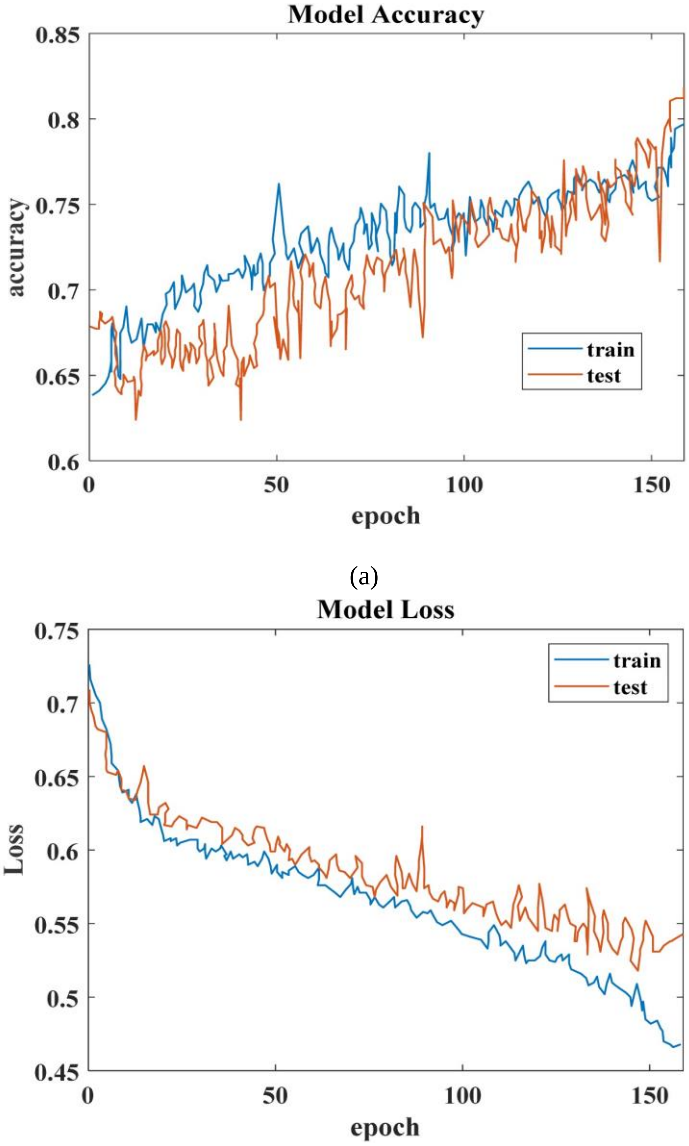


(a)

(b)

Figure 4. 8: (a) The accuracy graph of Bi-ENN for the train and validation samples on JAFFE and (b) the loss graph of Bi-ENN for the train and validation samples on JAFFE.

The Bi-ENN model achieved significantly low loss values for both the training and testing samples of the JAFFE dataset. Moreover, increasing the number of epochs substantially lowered the loss. Throughout 150 epochs, the neural network consistently maintains a loss value of around 0.5, which is significantly lower compared to numerous other neural networks. The model achieved a minimum error value using Adam optimization, which effectively identified the best weight value for training.

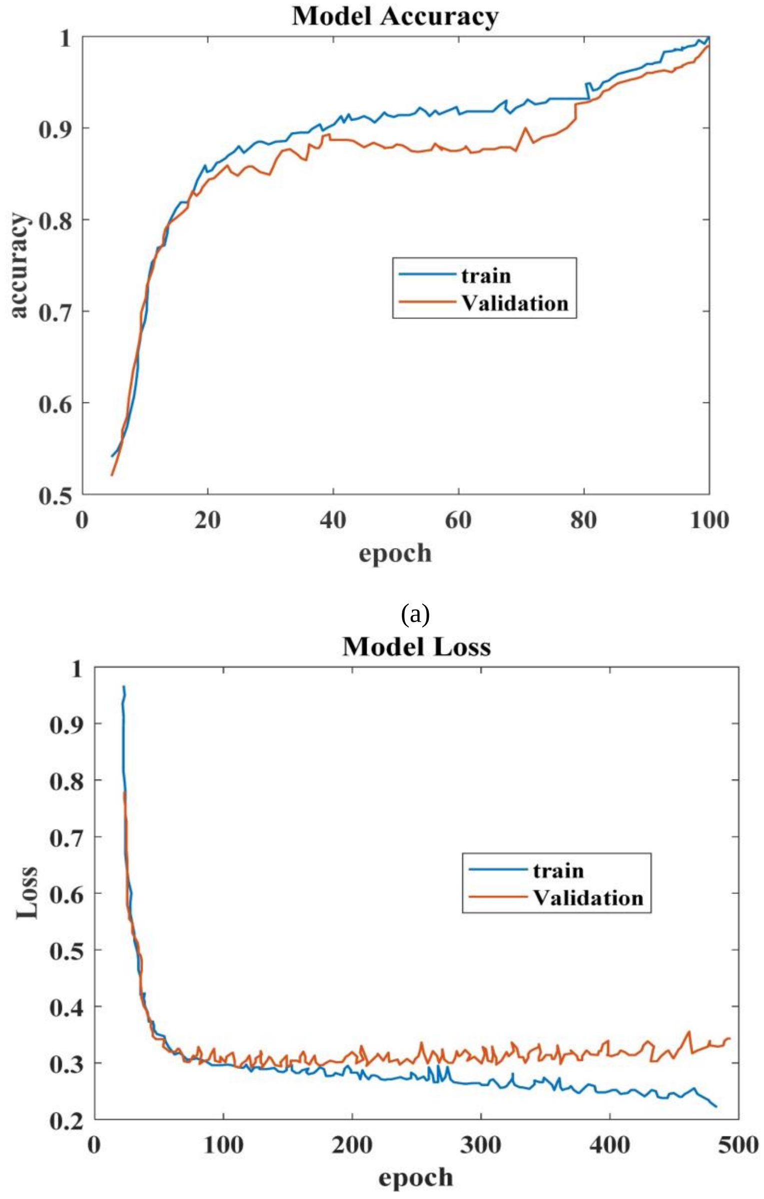


(a)

(b)

Figure 4. 9: (a) Bi-ENN's accuracy graph for the train and test samples on CK+ and (b) the loss graph for the train and test samples on CK+.

The Bi-ENN method uses pre-processed images from the CK+ dataset and extracts significant discriminative features for training. Between epochs 10 and 50, The scores of the training and testing samples are comparable. Following 100 epochs, the training and testing samples show comparable accuracy, showing that the model can be trained with varied input samples to enhance classification.



(b)

Figure 4. 10: (a) Bi-ENN's accuracy graph for the train and validation samples on CK+ and (b) the loss graph for the train and validation samples on CK+.

The Bi-ENN loss rates for both training and testing samples of the CK+ dataset are consistent, as the model yields nearly identical error values. The application of the Adam optimization algorithm within the hidden layers of the neural network decreases the error value. Upon completion of 500 epochs of training, the error value for the training samples is 0.2, whereas for the validation samples, it is 0.34. The proposed Bi-ENN shows a superior recognition rate compared to the existing Bi-LSTM and ENN models on the JAFFE and CK+ datasets, surpassing the performance of other evaluated neural networks. The Bi-LSTM model outperformed the ENN model, achieving better recognition rates. The Bi-ENN algorithm attained an overall recognition rate of 98.57% on the JAFFE dataset and 98.75% on the CK+ dataset. Thus, the proposed Bi-ENN enhances the performance and accuracy of the ENN by integrating both forward and backward trainin.

4.7.5 Findings of Image Filtering

The ECF filtering approach is utilized for image pre-processing, which is subsequently followed by the extraction of geometric and appearance-based characteristics. The most effective characteristics are subsequently chosen using EBRO and given to the Bi-ENN classifier for facial expression categorization. In the pre-processing step, images are loaded from datasets and subjected to noise and distortion removal techniques. Subsequent to the elimination of salt and pepper noise using DMF, the resultant image undergoes processing with AWF to eradicate additive noise. The Naturalness Image Quality Evaluator (NIQE) score assesses original and pre-processed images to determine the pre-processing stage.

The feature extraction step utilizes the Manhattan distance measure and a 2D log-Gabor filter to extract prominent facial characteristics. Changes in skin and texture are regarded as the extraction of features based on appearance. The suggested strategies extract a total of 71 facial features, utilizing the Manhattan distance calculation to enhance the information obtained from each facial point. Subsequently, The 2D log-Gabor filter extracts intensity and texture. This filter detects skin and texture variations to optimize feature extraction and facial expression recognition.

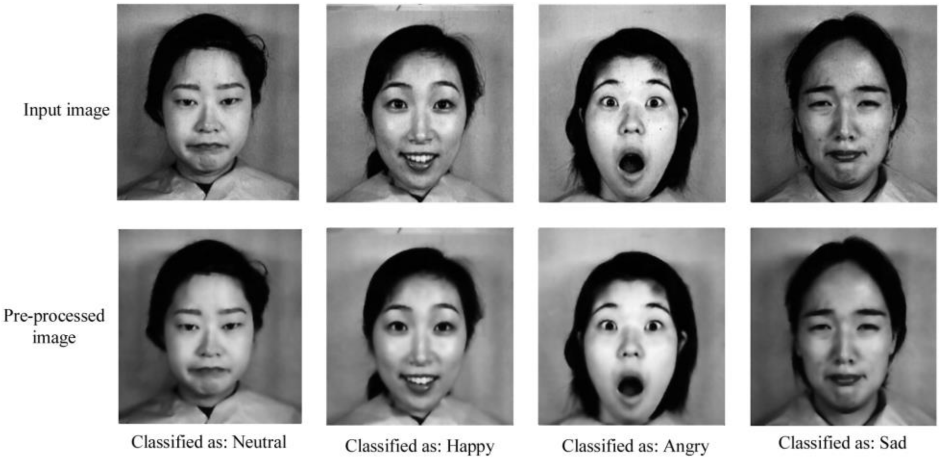
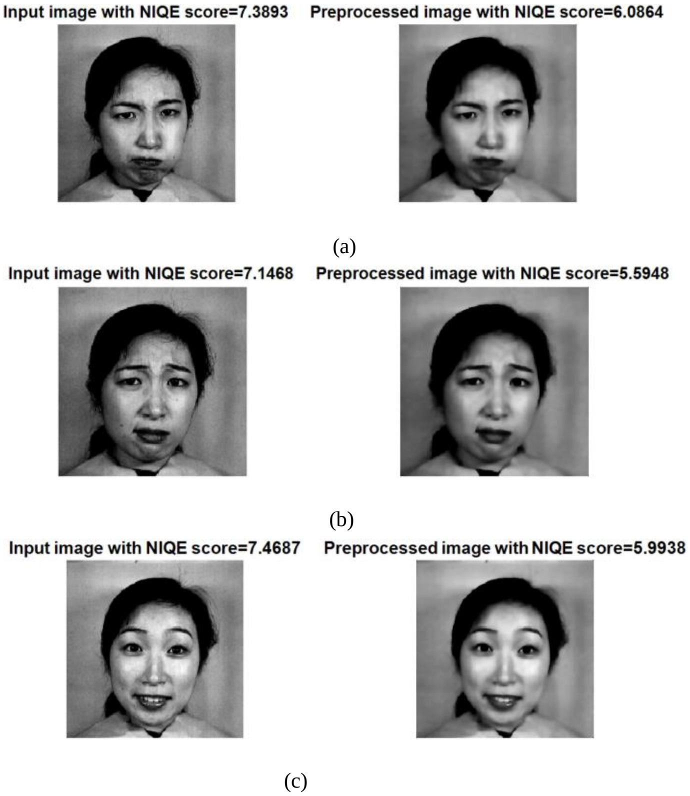


Figure 4. 11: Proposed pre-processing phase visualization.


Input image with NIQE score=7.3893
Preprocessed image with NIQE score=6.0864
Input image with NIQE score=7.1468
Preprocessed image with NIQE score=5.5948
Input image with NIQE score=7.4687
Preprocessed image with NIQE score=5.9938

(a)

(b)

(c)

Figure 4. 12: The NIQE data is shown visually. These are the NIQE scores for different facial expressions: (a) angry face, (b) disgusted face, (c) happy face.

The EBRO algorithm is used for the extracted feature set in order to identify the most distinctive facial traits. For the next step, a set of 29 highly discriminative features is chosen from each input image. These features play an essential part in accurately detecting facial emotions by capturing changes in the face. In conclusion, the proposed model effectively classifies facial emotions by removing noises and distortions, performing the extraction of geometric and appearance-based features, and subsequently choosing the most distinguishing features for classification.

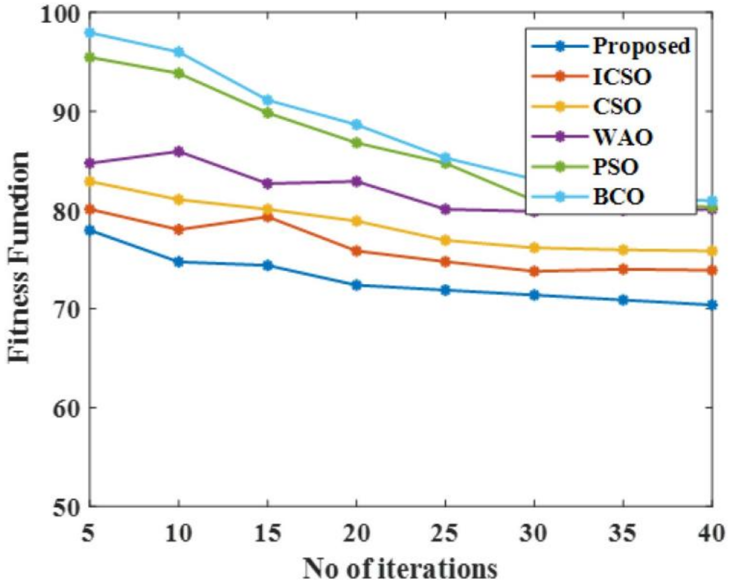
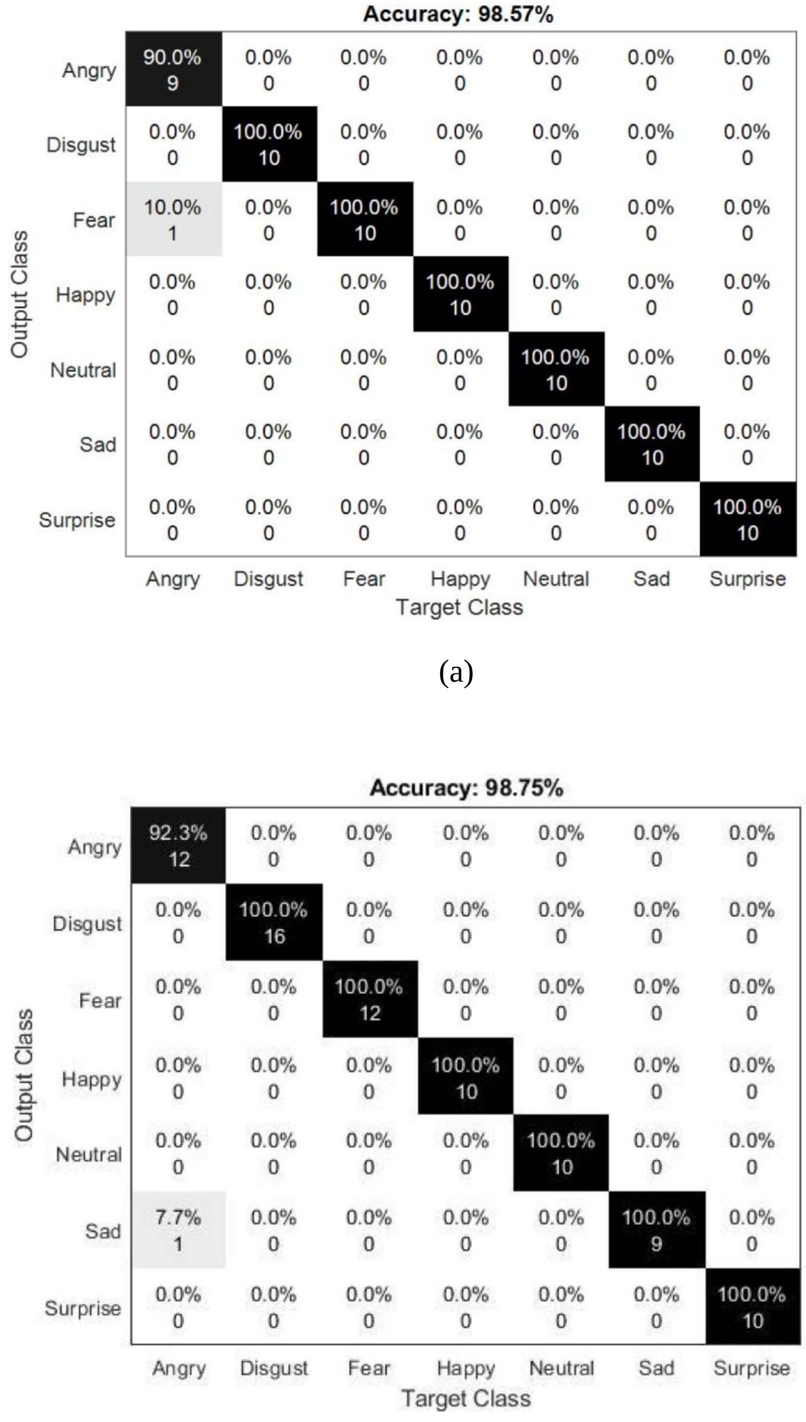


Figure 4. 13: Comparison of convergence graphs between the proposed EBRO algorithm and other algorithms.

The implementation starts after the selection of suitable characteristics for all photos. The model is trained in both forward and backward orientations to gather information from both past and future contexts. The proposed EBRO algorithm shows a convergence rate of 70% over 40 iterations, surpassing the performance of other algorithms documented in the literature, which was the ICSO algorithm, a refined iteration of CSO. The ICSO produced the most optimal outcomes when compared to other algorithms.



Accuracy: 98.57%

Output Class \ Target Class	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	90.0% 9	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Disgust	0.0% 0	100.0% 10	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Fear	10.0% 1	0.0% 0	100.0% 10	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Happy	0.0% 0	0.0% 0	0.0% 0	100.0% 10	0.0% 0	0.0% 0	0.0% 0
Neutral	0.0% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 10	0.0% 0	0.0% 0
Sad	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 10	0.0% 0
Surprise	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 10

(a)

Accuracy: 98.75%

Output Class \ Target Class	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	92.3% 12	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Disgust	0.0% 0	100.0% 16	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Fear	0.0% 0	0.0% 0	100.0% 12	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Happy	0.0% 0	0.0% 0	0.0% 0	100.0% 10	0.0% 0	0.0% 0	0.0% 0
Neutral	0.0% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 10	0.0% 0	0.0% 0
Sad	7.7% 1	0.0% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 9	0.0% 0
Surprise	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 10

(b)

Figure 4. 14: The confusion matrix of the suggested method for the (a) JAFFE dataset and the (b) CK+ dataset.

The confusion matrix for the proposed model's performance on JAFFE and CK+ datasets is presented in Figure 4. 11. Results show that the suggested model recognized facial emotions more accurately. The model performs nearly identically on both datasets. The seven facial emotions are classified further using Bi-ENN. The classifier labels each feature vector based on its unique features. During training, input feature vectors are supplied in both forward and backward directions for combining historical and future data. To facilitate classification, it improves the comprehension of the context of Bi-ENN. Figure 4.11 shows the proposed classification phase recognition rate on both datasets. The graph demonstrates that the suggested classifier achieved a higher level of accuracy in classifying facial emotions compared to the other classifiers being compared. The proposed classifier achieved an accuracy of 98.57% on the JAFFE dataset and 98.75% on the CK+ dataset. Overall model performance is determined using precision, recall, and F-measure. The model yielded improved classification results for the CK+ dataset, showing a reduction in false positives. The model achieved maximum class accuracies for CK+, except for the furious emotion, which had a slightly lower accuracy rate of 92.3%.

Both datasets show a ROC curve indicating that the true positive rate (TPR) surpasses the false positive rate (FPR). The model achieved an area under the curve (AUC) of 0.9875 on the JAFFE dataset and 0.9869 on the CK+ dataset. In the concluding chapter, the DBRO model achieved an AUC value of 0.989 for the JAFFE dataset, comparable to the proposed model.

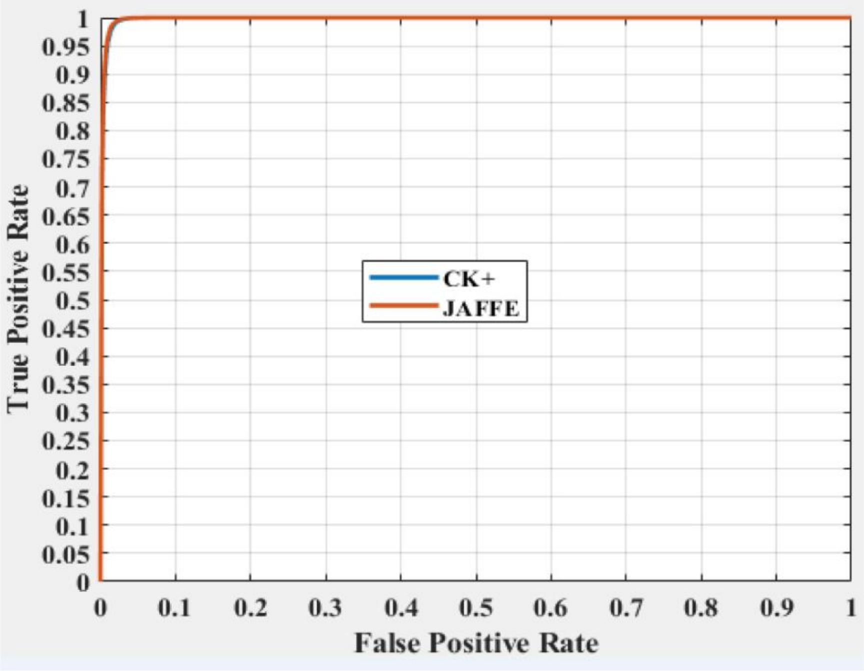


Figure 4. 15:Proposed model ROC curves on JAFFE and CK+ datasets.

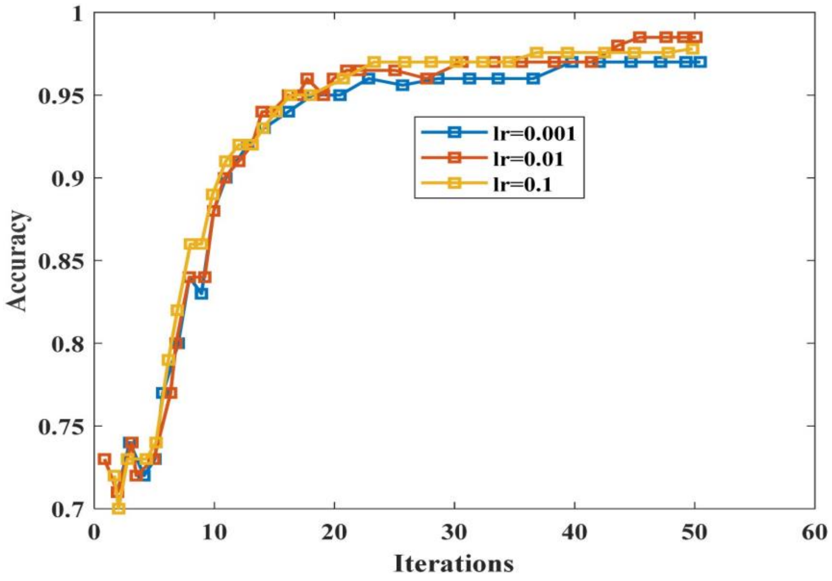


Figure 4. 16: Proposed JAFFE dataset accuracy by varying learning rates.

The selection of the learning rate for model training has an important effect on the overall performance of classification. Model precision on the JAFFE dataset at different learning rates is shown in Figure 4.13. The model was more accurate at 0.01 learning rate. The model's accuracy has been observed over varying numbers of iterations with respect to the learning rate.

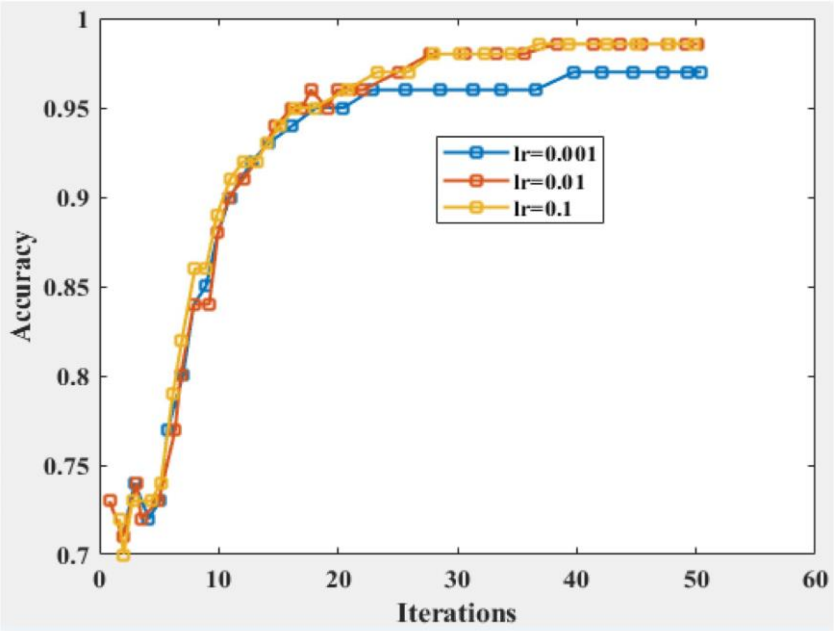


Figure 4. 17:Investigating the impact of different learning rates on the accuracy of the CK+ dataset.

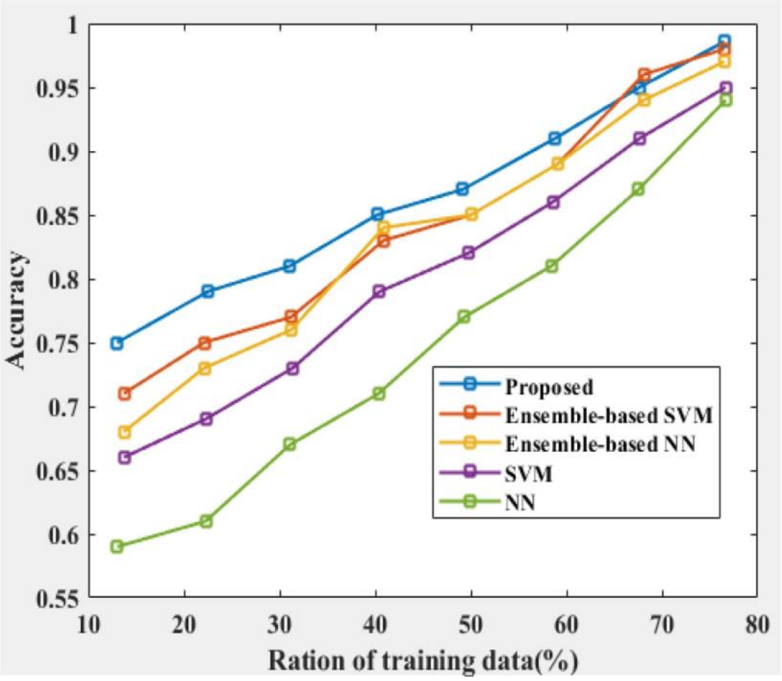


Figure 4. 18: Comparison of accuracy in relation to the percentage of training data in the JAFFE dataset.

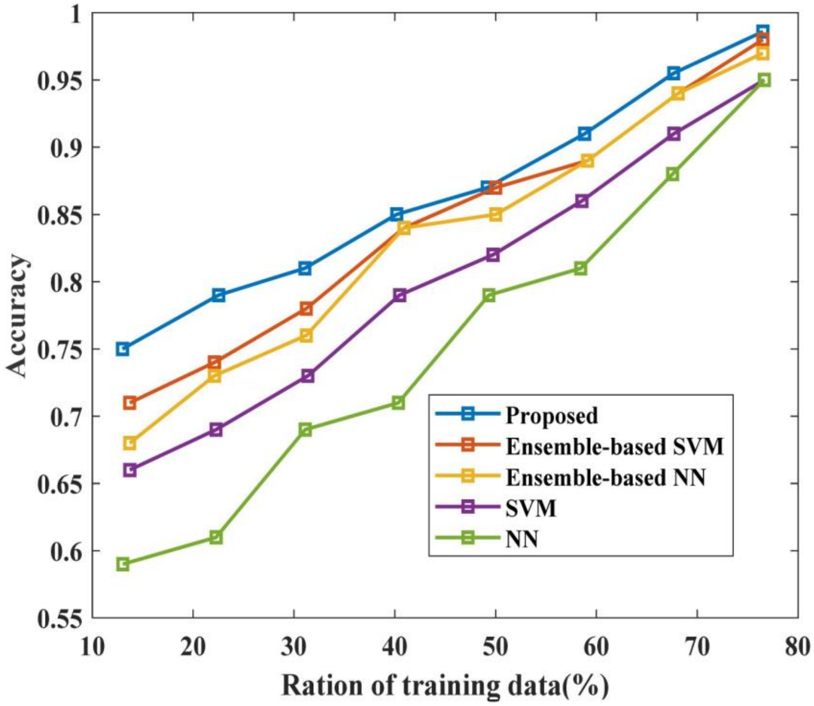


Figure 4. 19: Comparison of accuracy in relation to the percentage of training data in the CK+ dataset.

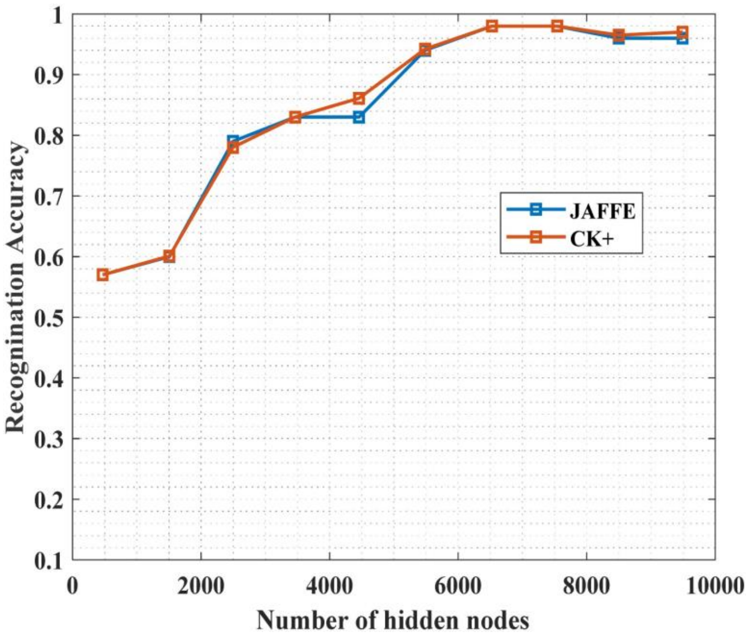


Figure 4. 20: Evaluating the precision of the proposed model by adjusting the number of hidden nodes on the JAFFE and CK+ datasets.

The overall accuracy of the training can be dependent on the total number of neurons in the hidden layers, as adjusting the number of hidden nodes can enhance the effectiveness of the training. The proposed framework had the maximum accuracy when the number of hidden nodes was increased. The model showed nearly identical performance on both the JAFFE and CK+ datasets, demonstrating its computational efficiency through its rapid convergence rate and exceptional identification accuracy.

4.7.6. Ablation Study

Ablation experiments were conducted on a proposed model to evaluate its performance and highlight the importance of each process. The model was divided into four modules: Modules 1, Modules 2, Module 3, and Module 4. The main reason for conducting ablation experiments is to identify the importance of each phase in the model. The findings showed that Module 1 had a higher degree of precision in terms of accuracy in comparison to the other modules. Module 2 demonstrated enhanced precision metrics but neglected the tuning procedure, leading to heightened error rates. Module 3 focuses on feature extraction, selection, and classification over pre-processing, lowering accuracy. Module 4 focuses exclusively on feature extraction and classification, skipping feature selection, and other stages. Controlling dimensionality concerns is essential to have a significant impact on the classifier's output, especially when input characteristics are not correctly provided [89].

Table 4. 1: Performance analysis of the proposed and existing classification techniques.

Dataset	Techniques	Precision (%)	Recall (%)	F-measure (%)	Specificity (%)
JAFFE	NN	94.75	94.15	91.26	94.12
	SVM	95.00	95.16	92.32	96.05
	Ensemble-based NN	97.55	97.23	93.10	97.06
	Ensemble-based SVM	98.63	98.25	94.12	98.20
	Proposed	98.57	98.70	98.57	99.77
CK+	NN	94.28	93.85	91.04	93.84
	SVM	94.72	94.85	91.82	95.57
	Ensemble-based NN	97.27	97.08	92.98	96.54
	Ensemble-based SVM	98.45	98.12	94.01	98.14
	Proposed	98.90	98.57	98.68	99.79

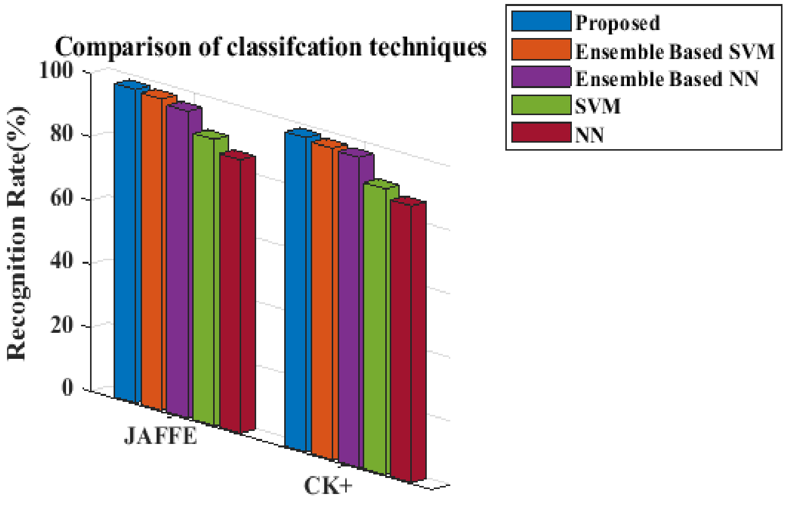


Figure 4. 21: Classification comparison for JAFFE and CK+.

The highest level of performance was attained in Module 1, which integrated all key stages of the proposed chapter.

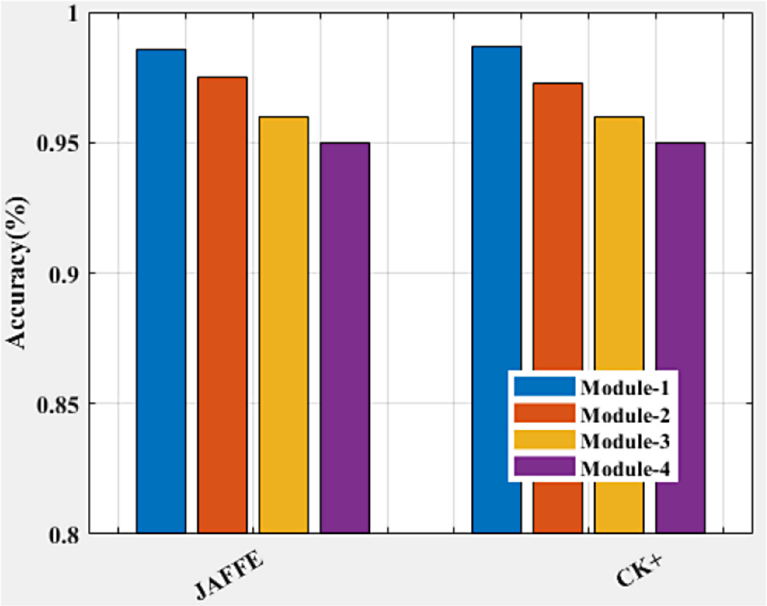


Figure 4. 22: Ablation study for JAFFE and CK+ dataset.

4.8 Summary

The Bi-ENN model, which is being recommended, is well-suited for classifying facial emotions due to its precise understanding of the distinguishing characteristics of the input elements. The model has undergone assessment using two distinct face emotion datasets, demonstrating its ability to effectively adjust to the input samples and its ease of training on various dataset types without compromising performance. The application of the Adam optimization method in the hidden layers has resulted in enhanced accuracy and reduced loss values for the model. This approach effectively reduces mistake rates in classification. The model also underwent evaluation using conventional ENN and Bi-LSTM algorithms to ascertain enhancements in performance. In classic ENN, the context layer only retains historical knowledge, which is diminished during the training phase. Forward and backward training are implemented to improve the available information for classification.

After 500 epochs of training, the model's accuracy showed substantial improvements at consistent intervals of 10 epochs. The loss curve shows underfitting issues in the early iterations, which are mitigated by augmenting the number of epochs. The reduction in loss value is ascribed to the implementation of the Adam optimization method for parameter tuning. The proposed Bi-ENN model is evaluated based on precision, recall, F-measure, accuracy, specificity, and recognition rate. It exceeds the comparative methods in all these metrics. The JAFFE dataset achieved a precision rate of 98.57%, whereas the CK+ dataset reached a precision rate of 98.75%.

This chapter improves classification with a novel Bi-ENN method. A Facial Emotion Recognition (FER) system uses the Bi-ENN to classify emotions for different people. The simulations showed that the proposed neural network classified more accurately than other algorithms. Since contextual data for classification improved through training, the model accurately identified emotions. On two datasets, the proposed neural network yields nearly identical results. Thus, data sample distribution and volume do not affect the proposed model. Given the need for efficient classification across domains, Bi-ENN may be the better choice for stable classification.

Our implementation was 98.57% accurate on JAFFE and 98.75% on CK+. Medicine is evolving and requires classifiers to identify diseases in individuals. Besides FER systems, the proposed Bi-ENN is widely used in mental health classification in medicine. With some training modifications, it can solve any classification problem. In summary, the suggested model is well-suited for precisely categorizing facial emotions.

Chapter 5: Hybrid Multimodal Emotion Recognition Framework

The recognition of emotion from a single modality is not always achievable, as individual modalities are influenced by various factors. In previous two chapters, we identified emotions from facial expressions from the images, although with high classification results. But when it comes to different modalities, The current models are unable to achieve optimal accuracy in precisely identifying individuals' expressions. This chapter introduces a novel hybrid multi-modal emotion recognition framework, InceptionV3DenseNet, aimed at improving recognition accuracy. Contextual features are initially extracted from various modalities, including video, audio, and text. The simulations are performed on the MATLAB platform and assessed using the MELD and RAVDESS datasets. The results demonstrated that the proposed model is greater in efficiency and accuracy compared to the other models, achieving an overall accuracy of 74.87% in MELD and 95.25% in RAVDESS.

5.1 Introduction

Emotion recognition is a highly interesting domain that has received increased attention due to its extensive applications. Identifying emotions from input data is complex, and using a specific modality is reliable. To enhance recognition performance, emotion can be elicited from diverse modalities. Common modalities, including auditory, visual, and textual cues, provide greater understanding of an individual's emotional condition. Recent literature on emotion classification indicates that deep learning-based approaches achieve higher accuracy, thereby reducing the misclassification rate. The latent space features could clarify the modalities and emotional states behind the cues. These features feed the emotion classifier. It makes feature differentiation difficult for the classifier, which could lead to misclassification. This chapter introduces a hybrid deep learning framework to overcome the above limitations.

This chapter introduces a deep learning-based multimodal emotion recognition framework that is both effective and efficient. In addition, a novel and efficient deep-learning neural network developed specifically for detecting facial emotions represents the primary contribution of the proposed research. The following are the primary contributions of this chapter's research work:

a. Recognizing emotions from several modalities is a highly complicated task that requires an extensive understanding of their properties. In this study, a novel emotion identification system employs three primary modalities: audio, video, and text.
b. To correctly detect the emotional signals, the feature is retrieved independently from all three modalities at first. The video extracts lighting, motion, and color features. The audio characteristics include measures such as the zero-crossing rate, Mel frequencies cepstral coefficient (MFCC), energy levels, and pitch information. Textual features comprise unigrams, bigrams, and TF-IDFs.
c. To improve classification performance, the proposed method integrates features before classification through multi-set integrated canonical correlation analysis (MICCA).
d. Implement the novel feature selection model, "Inception V3DenseNet architecture," which combines Inception V3 and DenseNet models.
e. Perform extensive assessments to utilize the various datasets to demonstrate the proposed model's performance efficiency compared to existing models.

5.2 Proposed Method

In this chapter, a multimodal emotion detection framework is developed to identify people's emotions correctly. The suggested conceptual structure includes input from the most relevant modalities, covering audio, video, and textual. In order to address the complex structure of the subject, the proposed methodology presents a pragmatic hybrid framework for precise emotion classification. The suggested model may be divided into three major phases, which are as follows:

a. Feature extraction
b. Feature fusion
c. Classification

Features from different methods are collected and integrated in phase 1 before they are put into the emotion recognition classification model. Figure 5.1 illustrates the structure of the proposed framework.

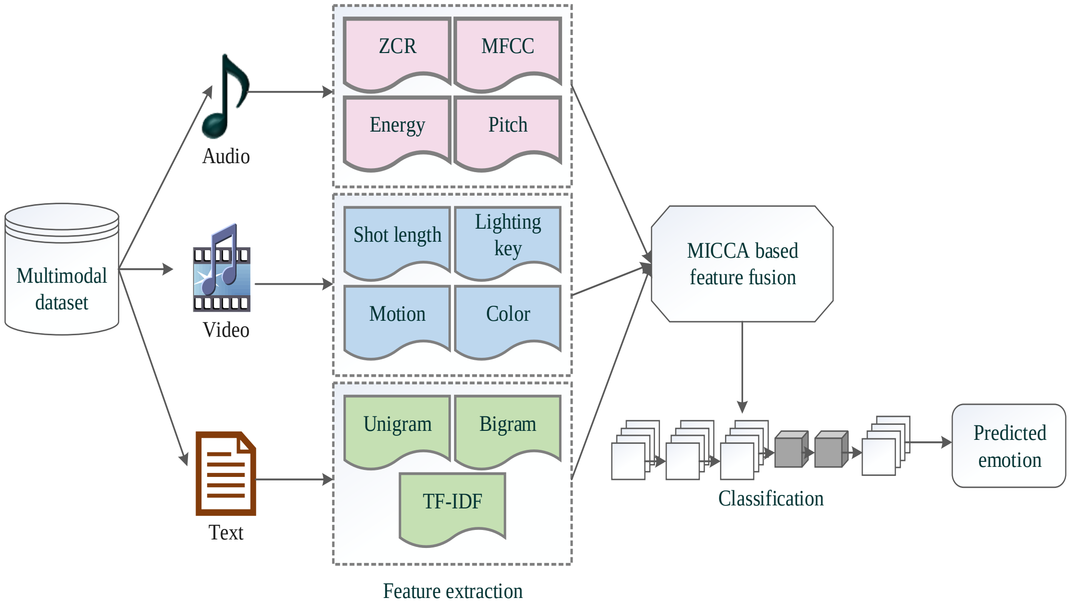


Figure 5. 1: The proposed framework.

The proposed framework starts by extracting latent space feature vectors. The video input provides short duration, key lighting, motion, and color attributes. Audio features such as zero crossing rate, cepstral coefficients, energy levels, and pitch are determined from audio signals in the analysis of the audio domain. Extraction of TF-IDF, unigrams, and bigrams from user comments in textual input enables concurrent analysis and processing. The feature is built using a probability-based method at the feature level. The feature that is obtained is incorporated into the hybrid classification model used for predicting emotions. To improve prediction accuracy, the proposed hybrid model integrates the numerical architectures of InceptionV3 and DenseNet. An attention mechanism has been added at the end to enhance performance. The weight-adjusting method used in the meta-heuristic-based honey badger algorithm increases the classifier's overall performance (HBA).

5.2.1 Phase 1: Feature Extraction

One of the crucial steps in the suggested approach is feature extraction, in which the feature is retrieved separately from the considered modalities. Three modes, such as auditory, visual, and textual, are examined in the study. The following are the primary characteristics taken from each input modality:

a. Audio Feature Extraction

Audio is a key modality for recognizing people's emotions. Audio feature extraction is crucial in the research, as the classification process relies heavily on these characteristics. The procedure includes quantifying significant audio characteristics such as zero crossing rate, Mel frequency cepstral coefficient (MFCC), energy, and pitch. Audio feature extraction plays a significant role in audio processing, music information retrieval, audio effect creation, and audio synthesis. Audio features are frequently used in design, evaluation, and analysis. However, a wide range of feature extraction tools are available to the community. Audio feature extraction is an essential component of contemporary research and advancement in audio signal processing. Therefore, numerous libraries and toolboxes for collecting audio features have been developed. Some tools are designed for managing workflows, pre-processing, and batch operations, whereas others prioritize algorithmic efficiency, parallelization, specific programming environments, or platforms. While there has been significant progress and research in audio signal processing and feature extraction, research on evaluating and selecting appropriate feature extraction tools and their respective applications is absent [90]. Here is a description and mathematical representation of the feature:

i. Zero-crossing Rates

The zero-crossing rate refers to the frequency at which a signal intersects the zero line on the x-axis or the rate at which the signal changes per unit of time. The zero crossing rates can be mathematically expressed as:

$$ZC_h = \frac{1}{2N}\sum_{q=1}^{N}\left|sign(x_h(q)) - sign(x_h(q-1))\right| \tag{5.1}$$

$$sign(x) = \begin{cases} 1, & x > 0 \\ 0, & x \le 0 \end{cases} \tag{5.2}$$

where, $x(h)$ The audio input signal is segmented using a sliding window, $x_h(q)$ indicating the h^{th} segment's time sequence $h \in [h_1, h_2]$.

ii. MFCC Coefficient

MFCC’s capture the spectral features of an input signal using a non-linear Mel scale. This method utilizes the technique to examine the cepstral coefficients. The goal of employing MFCC for hand gesture detection is to investigate the MFCC's utility for image processing. It has previously been utilized in voice recognition and speaker identification. Constant window size and constant shift are employed to put audio signals into frames.

For each frame we need to the followings:

- Calculate the "Fast Fourier Transform" (FFT)
- The Mel scale in relation to pack frequencies
- Calculate the logarithm
- Calculate the discrete cosine transform (DCT)

FFT: The FFT method has two variants, both comparable in computing complexity. Both methods permute the order of the input and output row-column indices.

Pack Frequencies with respect to the Mel scale: The Mel scale is a pitch scale in which listeners see equal distances between pitches. The scale is based on a 1000 Hz tone that is 40 dB higher than the audience's threshold and has a pitch of 1000 meals. Lower frequencies are better at recognizing changes than higher frequencies. It can differentiate between 500 and 1000 Hz but has difficulty differentiating between 10,000 and 10,500 Hz despite a comparable distance between the two pairs [91].

iii. Energy

The overall strength of the signal or the amount of energy in the signal aids in obtaining additional information about the emotion contained in the signal. An audio signal will have high energy if it conveys genuine excitement. The total energy in the signal can be determined using the formula mentioned below.

$$E_n = \sqrt{\frac{1}{N}\sum_{q=1}^{N}\left(x_h(q)^2\right)} \tag{5.3}$$

iv. Pitch

The pitch of a signal is typically higher for intense or stimulating signals and lower for weak signals. The signal's pitch is usually identified using autocorrelation analysis. Calculations can be conducted within the framework of a sinusoidal random process function $x[q]$:

$$x[q] = \cos(\omega_0 q + \Phi) \tag{5.4}$$

The autocorrelation $x[q]$ can be provided as follows:

$$A[q] = \varepsilon\{x*[q]x[q+h]\} = \frac{1}{2}\cos(\omega_0 h) \tag{5.5}$$

Pitch of the signal is determined using the autocorrelation peak. An estimate $A[q]$ can be calculated as follows:

$$\hat{A}[q] = \frac{1}{\xi}\sum_{s=0}^{\xi-|h|}\left(\upsilon[s]x[s]\upsilon[s+|h|]x[s+|h|]\right) \tag{5.6}$$

Where, $\upsilon[s]$ is the length of the window and the anticipated value for $\hat{A}[q]$ is determined using the following formula:

$$\varepsilon\left\{\hat{A}[q]\right\} = \left(1 - \frac{|h|}{\xi}\right)\frac{\cos(\omega_0 s)}{2}; \quad |h| < \xi \tag{5.7}$$

b. Visual Feature Extraction

The video's characteristics depend on the input video. The technique suggested includes extracting video properties like shot duration, lighting intensity, motion, and color. Below is an in-depth description and mathematical depiction of the characteristics:

i. Shot Length

A shot in video production is a sequence of frames captured by the camera with consistent coloration across consecutive images. Color histogram (CH) methods enhance the precision of identifying shot durations in a video. Typically, video frames are shown in the HSV colour space in order to determine the duration of each shot. Chroma values are computed for the hue, saturation, and value channels of each video frame. It represents the feature matrix for the frames captured at various time intervals:

$$\chi^h = \begin{bmatrix} x^h \\ x^{h-1} \\ \cdots \\ \cdots \\ x^{h-\ell+1} \end{bmatrix} \tag{5.8}$$

where, $h = \ell, ...N$, ℓ denotes the length of the window and N denotes the total number of frames following this, the matrix is decomposed using the singular value decomposition (SVD) method χ^h . $e_1, e_2, ...e_\ell$ Are the Eigen values, and e_1 the maximum value is shown. The matrix's rank is $\Re^h$ computed and χ^h may be expressed as the total number of e_q with $\frac{e_q}{e_1} > \tau$, where τ is a threshold value. The calculated rank has two essential properties: 1) The image contents in the present and previous frames vary if $\Re^h > \Re^{(h-1)}$ 2) The visual content of the present frame remains constant and seamlessly replaces the preceding shot if $\Re^t < \Re^{(t-1)}$. The frame with the highest rank is referred to as the commencement frame of a shot. The frame with the highest rank is known as the commencement frame of a shot. The shot length for $\Re^h > \Re^{(h-1)}$ and $\Re^h > \Re^{(h-1)}$ is calculated by performing a difference operation on the two frames in question.

ii. Lighting Key

The lighting key is an essential visual element that represents the light quality present in the video. Two distinct lighting techniques can be used to obtain the primary characteristics of the lighting:

low-key and high key. Since the contrast in the high-key scene is relatively low, it differentiates it from the other scenes. Furthermore, the contrast between the brightest and darkest light in the high-key scene is negligible; high-key sequences are typically linked to amicable scenes.

Three-point lighting is a conventional way of illuminating a subject in a scene by shining light from three different positions. Essential lighting, fill lighting, and backlighting are the three forms of lighting. Lighting keys explain how a scene will be lit and reflect the emotional story of the scene, whereas storyboards tell the narrative tale. A good lighting key originates from the heart and needs empathy and artistic thought. The key light is the primary light source in a movie or photograph. High-key lighting leads to brightly illuminated subjects, with more filled light and less harsh shadows. Fill lights provide additional light to a scene, reducing the contrast. Lighting plays an important role in video and film production because cameras react to light differently than the human eye. Cameras cannot capture complex information and lighting changes the human eye can detect.

The scene is defined by subdued lighting, creating a notable contrast with a dark background. The black background is often associated with frightening or intense imagery because of its unappealing nature. Scene lighting is directly correlated with the pixel brightness in the image. High-light scenes are characterized by a greater number of pixels with high-light values, whereas low-light situations predominantly consist of pixels with lower light values. The illumination key for a frame, Γ, can be defined as follows:

$$\kappa_{\Gamma} = \alpha_{\Gamma} \times \sigma_{\Gamma} \tag{5.9}$$

Where, κ_{Γ} indicates the lighting key, α_{Γ} is the average, and σ_{Γ} is the variance of the frame value.

iii. Motion Feature

The temporal information given by the motion characteristics in a video enables the prediction model to determine emotions effectively. The proposed work extracts motion data from the input video using block-matching techniques. The main objective of developing this method was to identify consecutive segments within a video stream. The notion that objects are generated by the sequential movement of background patterns and objects within a video frame is embodied in this program. After partitioning the present frame into macro chunks, a comparison is made with the preceding frames. An animation vector is generated to depict the movement of a macroblock. The motion vector is computed for every individual frame block.

iv. Color Feature

Color is essential in videos to display content and plays a crucial role in evoking emotions, especially during emotional scenes. The video's color is strongly linked with its color energy. The following is the mathematical formula for color energy:

$$C_E^r = \sum_{k=1}^{P_T} \frac{\varsigma_k \times \nu_r}{\sigma_H \times P_T} \tag{5.10}$$

where, r represents the frame, ς_k *and* ν_r are the pixel's value and saturation k, σ_H represents the hue histogram's standard deviation and P_T is the total amount of pixels in the specified frame.

c. Textual Features Extraction

Textual features are employed to generate specific emotions from written material. The proposed work's textual features originate from the textual comments provided by the viewing audience. Multimodal textual features are employed to analyze the emotional content of texts, with the unigram, bigram, and TF-IDF features being the most prevalent. The bigram feature analyses the sequential occurrence of two words throughout the entire document, as opposed to the unigram feature, which only requires a single word to be present. In a document, the TF-IDF determines the significance of each word. The concept includes two distinct components: inverse document frequency and term frequency. The inverse document frequency metric measures the semantic content of individual words within a specific set of documents. Term frequency estimates the frequency of a specific term within an individual document. The following are the mathematical formulations:

$$hf - qdf(\beta,\gamma,\phi) = hf(\beta,\gamma)\cdot qdf(\beta,\phi) \tag{5.11}$$

$$hf(\beta,\gamma) = f_c(\beta,\gamma) \tag{5.12}$$

$$qdf(\beta,\phi) = \log \frac{|\phi|}{1+|\gamma \in \phi : \beta \in \gamma|} \tag{5.13}$$

Where, β denotes the word, γ denotes the document, ϕ denotes the corpus, $f_c(\beta,\phi)$ denotes a raw count of words β in the document, γ and $|\gamma \in \phi : \beta \in \gamma|$ denotes the document number with $hf(\beta,\gamma) > 0$. All of these elements are taken from the textual content to determine emotions.

5.2.2 Phase 2: Feature Fusion

The previously obtained features are put together to form the input for the classifier. The proposed methods use multiset canonical analysis (MCCA) to combine data from three senses. Canonical Correlation Analysis (CCA) is a popular statistical technique to uncover and assess the connection between two variables. The conventional CCA is restricted to fusing only two feature vectors. The MCCA is a CCA extension that can conduct feature-level fusion on multiple feature vectors. The multiset integrated CCA (MICCA) approach is used in the study to achieve an improved fusion of derived characteristics.

Consider the set of characteristics that will be combined as g. The feature group correlation criteria function can be defined as follows.

$$\Upsilon = \left(\lambda^{(1)}, \lambda^{(2)}, \ldots.\lambda^{(g)}\right) = \sqrt{1-\det\left(G\left(\psi^{(1)}, \psi^{(2)}, \ldots.\psi^{(n)}\right)\right) \Big/ \prod_{n=1}^{g} \left\|\psi^{(n)}\right\|} \tag{5.14}$$

Where, $\lambda^{(g)}$ denotes the feature projection g and G is the gram matrix with a vector ψ. The proposed work utilizes feature fusion by combining feature vectors using the summation approach. The input characteristics are combined during feature fusion to increase their discriminative power for classification.

5.2.3 Phase 3: Classification

The network model under evaluation utilizes the fused feature from the feature fusion step to generate appropriate labels for the input feature. The paper suggests the use of the Inception V3 DenseNet, a hybrid architecture incorporating deep feature learning capabilities, for emotion classification. The Inception V3 model works together with the DenseNet architecture to optimize classification performance, as implied by its name. Both algorithms are variations of the CNN model, widely recognized as the leading model for emotion classification. Furthermore, DenseNet exhibit far more intricate architectures. It implements a feed-forward technique by connecting each layer to every succeeding layer. The proposed model has the potential to address the vanishing-gradient problem, enhance the propagation of features, encourage the reuse of features, and substantially decrease the number of parameters.

Classification of images is possible with the Inception V3DenseNet model using training data and derived discriminative attributes. This model comprises multiple convolution layers to extract features using kernels of varying sizes. Adding thicker blocks increased the density of the previous Inception V3 model. The structure of the proposed Inception V3 DenseNet model is illustrated in Figure 5.2.

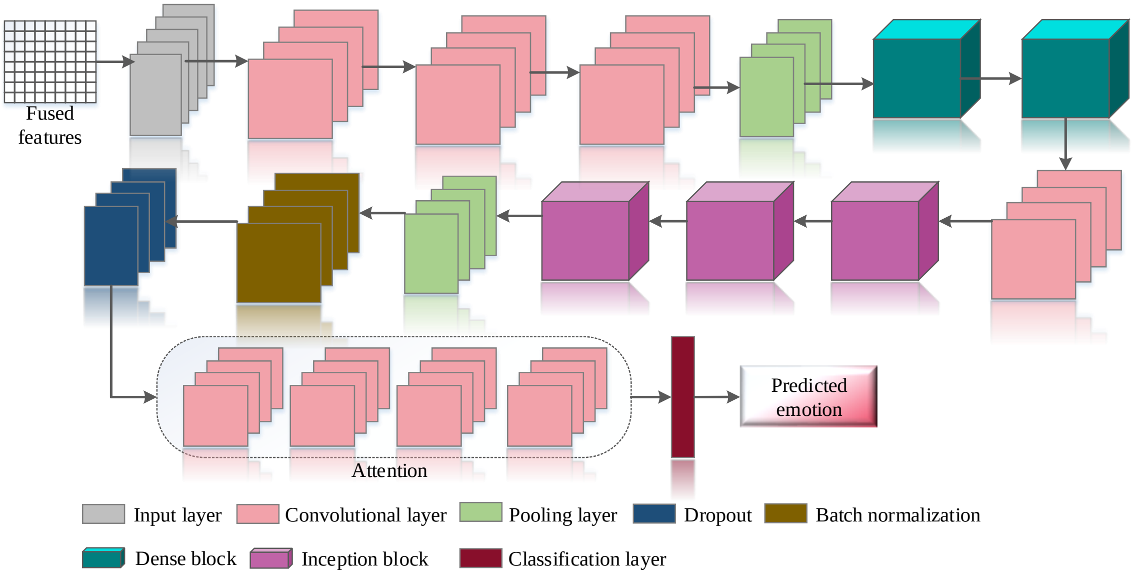


Figure 5. 2: The proposed Inception V3 DenseNet model's structure for multimodal emotion identification.

The image depicts the proposed model as a sequence of convolutions applying kernel sizes that vary. The input features pass one pooling layer and three convolutional layers before going forward. A single convolution operation is performed after adding two dense blocks into the model. The features go through three inception blocks, one batch standardization layer, and one dropout layer. The result obtained from these layers is then processed through the convolutional layer to enhance the categorization process. The classification layers are the network's last layers, where the input characteristics are labeled with distinct emotions. The processes and mathematical formulations that occur within the layers are as follows:

a. Convolutional layer

The convolutional layer has been intentionally engineered to conceal visual content and acquire adaptive knowledge of manipulation detection qualities. It demonstrates that the proposed method can independently discover how to recognize numerous image modifications independent of preprocessing or preselected characteristics through a series of trials. While CNNs can acquire acceptable features for object identification through change, their current state does not make them suitable for detecting image alterations. Convolutional layers are used to extract features that accurately represent the content of an image, as opposed to learning filters that identify traces caused by editing and manipulation. A convolutional layer with numerous convolutional filters applied in parallel with the image constitutes the initial layer. Typically, a pooling layer is appended after a convolutional layer to reduce the dimensionality of the feature maps. It reduces the computational expense related to network training and mitigates the probability of over fitting. Pooling layers divide feature maps into minuscule windows, which may occasionally overlap and retain a single value within each window. Maximum pooling and mean pooling are the most prevalent forms; they retain each window's maximum and mean values, respectively.

The convolutional layer is a key element in the classification model, extracting the vital characteristics needed for precise labeling. The mathematical theory that controls feature extraction by the convolution layer is as follows:

$$c_a^o = f\left(\sum c_a^{o-1} * r_{va}^o + b_a^o\right) \tag{5.15}$$

Where, f is the non-linear function, r_{va}^o is the layer's convolutional kernel o^{th} with input and output maps, a and b_a^o represents the associated bias vector. The convolutional kernel iteratively analyses the input to extract essential features. Training the network model with the extracted features is an important phase in the suggested approach.

b. Pooling layer

Pooling layers decrease the dimensions of the feature maps. Therefore, it reduces the number of parameters to be learned and the computational workload within the network. The pooling layer condenses the characteristics within a specific region of the feature map generated by a convolution layer. Pooling is a technique used in convolutional neural networks to apply the features extracted by convolutional filters and help the network identify features regardless of their location in the image. Poll the network for information from devices you may use to monitor the devices' activities. The Network Manager polls the network by sending queries to network devices regularly. The feature pooling layer aims to remove duplicate information while preserving more important information. It is critical for any feature pooling approach to determine which characteristics are informative and which are irrelevant/redundant [92].

The pooling layers are the next layer in the network, and their primary function is to lower network parameters and computational complexity. Furthermore, this layer is capable of managing the network's over-fitting issues. The following is the mathematical formulation for the pooling layers:

$$c_a^o = f\left(\tau_a^o \, pooling\left(c_a^{o-1}\right) + b_a^o\right) \tag{5.16}$$

Where, τ_y^k is the pooling procedure and $pooling(\cdot)$ is the training parameter.

c. Dense block

A Dense Block is a convolutional neural network module that directly connects all layers (with matching feature-map sizes). It was first proposed as a component of the DenseNet design. In the proposed network model, dense blocks enhance classification performance while decreasing the total number of network parameters. The dense block of DenseNet-BC is used in the suggested model since it gives better performance results. All known search-based approaches for dense-block identification in tensors are predicated on the assumption that tensors are small enough to fit in memory. The rectified linear unit (ReLU) serves as the activation function throughout the dense block [93]. The construction of a singular dense block in the proposed work is illustrated in Figure 5.3.

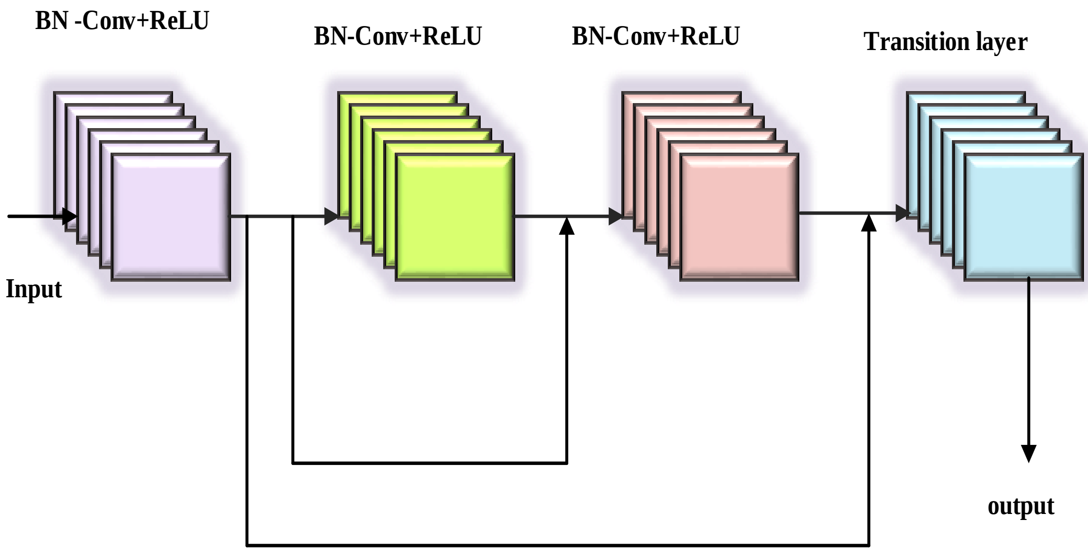


Figure 5.3: The proposed work uses the structure of a single dense block.

Following the analysis of three bottleneck convolutional layers is one transition layer per dense block. The transition layer consolidates the feature maps created by different layers inside the dense block once the aggregate features are transferred to the next network layer for training.

d. Inception block

Within a Convolutional Neural Network (CNN), an Inception Module is a component of the image model that imitates an ideal local sparse structure. The use of several filter sizes within a single image block is feasible, so overcoming the constraint of a singular filter size. Next, these filters are merged and sent to the following layer. An Inception network is a type of deep neural network that has an architectural structure made up of repeating components known as Inception modules. As previously stated, the technical intricacies of the inception module are the emphasis of this article.

Within the proposed model, the inception block exhibits competence in effectively managing filters of different sizes. Concatenation is the process of transferring feature maps from different layers with varying filter widths to the subsequent layer within the inception block. The inception block configuration in the proposed model is illustrated in Figure 5.4.

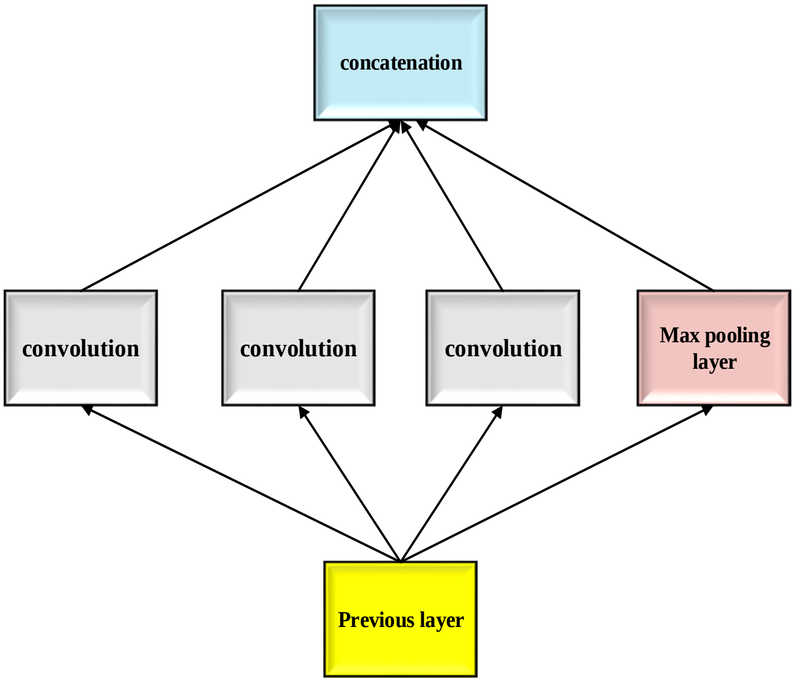


Figure 5. 4: The inception block's structure.

The proposed model's initial configuration includes three convolutional layers and one max pooling layer. The outputs from these layers are combined and then sent to the subsequent layer in the network. Inception block layers contain convolutional layers succeeded by a max pooling layer.

e. Attention mechanism

The attention mechanism allows the decoder to target key elements of the input sequence by using the weighted total of the encoded input vectors, with the most relevant vectors being assigned the highest weights. The attention mechanism enhances the efficiency and accuracy of perceptual information processing significantly. Most deep learning attention mechanisms are developed for specific activities; therefore, they are mainly focused attention. Except for specific claims, the attention mechanism outlined in this work refers typically to focused attention in the neural network area classified images using the attention mechanism on the recurrent neural network model. The attention mechanism allows machine translation tasks to be translated and aligned at the same time. As a result, the attention mechanism has become a more prevalent component of neural networks and has been used for various tasks, including image captioning.

An attention mechanism is included after the network model to prioritize the essential aspects. It assists the model in correctly labeling the input characteristics. The proposed model's attention mechanism includes four convolutional layers that filter and select features according to their discriminative characteristics. This attention mechanism directs its output to the classification layer, which categorizes the input features into separate emotional states.

f. Weight tuning mechanism

The network models' weight values are chosen randomly, resulting in prediction errors. To avoid this, it is critical to fine-tune the weight in each cycle. This part offers a weight-adjusting method based on the HBA algorithm. The Honey Badger Algorithm (HBA) is derived from the foraging behaviors observed in honey badgers. The algorithm includes two main stages: excavation and honey extraction. The solution space is initially filled with a set of chosen weights. Specify the coordinates of the weights in the solution space:

$$w_q = lb_q + rand_1 \times \left(ub_q - lb_q\right) \tag{5.17}$$

Where, $ub_q \, and \, lb_q$ are the search space's upper and lower bounds, and $rand_q$ is a random number from 0 and 1. An intensity factor measures the distance between the honey badger and the optimal solution. The inverse square law is used for this calculation. The formula adjusts the density factor to transition from exploration to exploitation:

$$\partial = C_1 \times \exp\left(\frac{-t}{t_{\max}}\right) \tag{5.18}$$

Where, C_1 is a constant that $t_{\max}$ represents the maximum number of iterations. By introducing a flag, one can navigate the search space to find better solutions and prevent getting stuck in local optima. The location is adapted based on the excavation and honey production stages. The honey badger's activity resembles a cardioid motion during the digging phase. The honey badger tracks the honeyguide bird to locate honey during the honey season. Here is the formulation for updating the position.:

$$w_{newp} = w_{psol} + F \times ra \times \partial \times dist_q \tag{5.19}$$

Where, w_{newp} is the new position, w_{psol} is the solution position, F is the flag to change the search space, ra is a random value between 0 and 1, and $dist$ the distance information.

5.3 Experimentation and Result Analysis

The evaluation of the model's performance uses two distinct datasets: the Multimodal Emotion Lines Dataset (MELD) and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). The MELD dataset represents an enhanced version of the Emotion Lines dataset, incorporating data across audio, visual, and textual modalities. The dataset consists of identical dialogue sequences derived from the Emotion Lines collection. The MELD dataset consists of more than 1,400 conversations, amounting to 13,000 utterances, derived from the Friends television series. The dialogue includes several speakers, with each statement categorized into one of seven emotions: anger, disgust, joy, surprise, fear, neutral, and sadness.

Furthermore, the dataset contains sentiment annotations for each utterance, including positive, neutral, and negative labels. The RAVDESS dataset includes 7,356 files, a total of 24.8 GB in size. The dataset comprises audio recordings of two lexically equivalent statements presented in a neutral North American accent. A total of 12 female and 12 male professional actors provide the vocal performances. The conversation can be classified according to various emotions, such as joy, tranquility, sorrow, dread, repulsion, astonishment, and anger. The dataset includes songs categorized into five specific emotions: happiness, calmness, fear, sadness, and anger. Each emotion is characterized by two distinct levels of intensity: strong and normal, along with an additional neutral expression.

5.3.1 Simulation

The MATLAB platform is used for all simulations in this work. There are 26 layers in the proposed model. The hyperparameter values for the suggested model are shown below: This study uses the HBA tuning algorithm with 15 convolution layers, 4 pooling layers, 2 normalization layers, 40 batch sizes, 32 input sizes, an initial learning rate of 0.1, a dropout factor of 1, 100 initial populations, 10 maximum iterations, and 100 maximum epochs.

5.3.2 Performance Analysis

We have evaluated the proposed model's performance using two distinct datasets, and this section provides a comprehensive analysis.

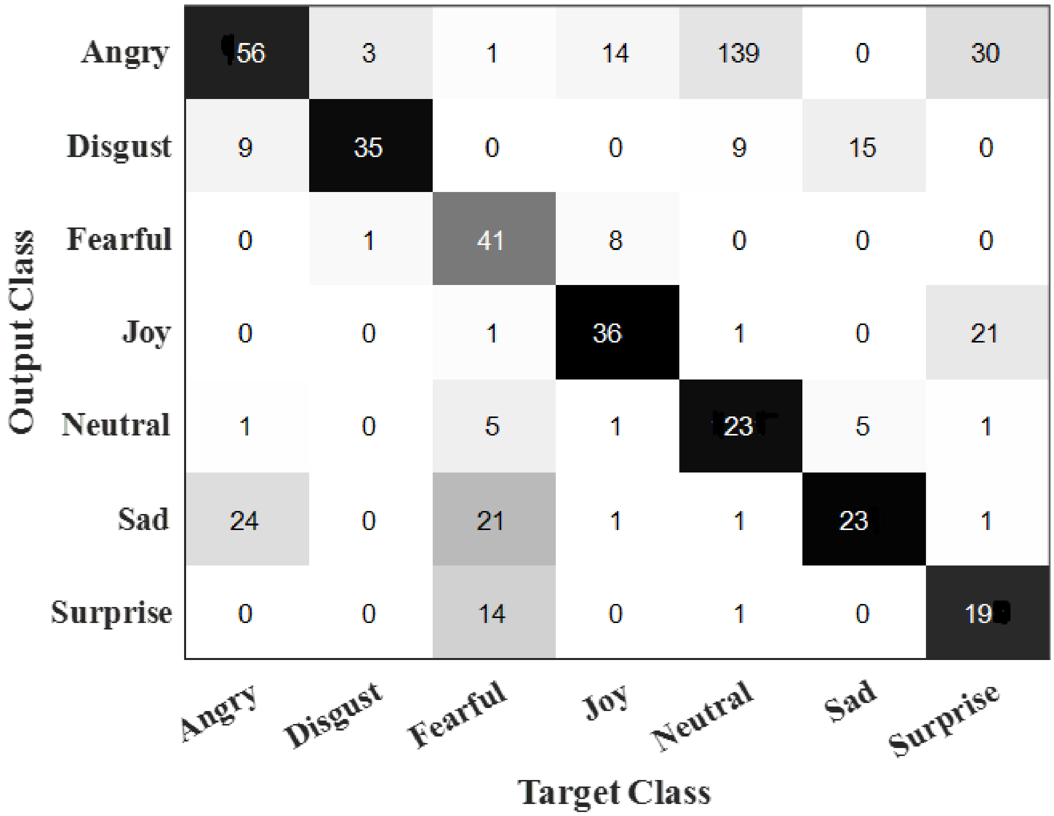
a. **MELD Dataset**

The proposed model's initial training is performed on the MELD dataset. To demonstrate its efficacy, after that, the results are looked at and compared to those from other models. The needed features are taken out of the MELD dataset data that is collected. Subsequently, these features are integrated to provide the classifier with a foundational basis for operation. After completion of training, the proposed classification model accurately categorized the input features. The subsequent results indicate that the proposed model demonstrates superior performance on the MELD dataset.

In Table 5, you can see how well the proposed and existing models did on the MELD dataset. It can be seen in the table that the estimated model got higher precision coefficients than the other methods. For each class, the performance metrics have been calculated on their own. In every class, the proposed model is better than the existing models in terms of precision, recall, and F-score. Among all the models that were compared, the cross-modality model produced better outcomes compared to the MMFA model. Furthermore, it has been noted that there is a significant enhancement in the performance metrics of the designed model. The model's remarkable ability to accurately label the input features can be due to its high discriminative capacity. Furthermore, the HBA algorithm significantly decreased the error rate by choosing the optimal parameter for successful classification. In correctly predicting emotions, the proposed model showed greater accuracy compared to the existing models.

Table 5. 1: Performance results of the proposed model on the MELD dataset

Metrics	**Classes**	**MMFA**	**Cross modality fusion**	**Proposed**
Precision	Angry	50.00	51.40	79.24
	Disgust	30.00	47.10	86.87
	Fear	50.00	50.00	46.53
	Joy	53.72	56.70	90.98
	Neutral	74.46	74.00	86.20
	Surprise	49.32	54.4	75.59
	Sad	40.37	49.60	89.16
Recall	Angry	37.50	44.7	77.00
	Disgust	14.00	11.80	47.99
	Fear	20.00	6.00	78.52
	Joy	60.79	57.20	90.61
	Neutral	83.31	84.90	95.47
	Surprise	67.41	61.40	89.31
	Sad	21.57	29.30	79.31
F-score	Angry	42.86	47.80	77.67
	Disgust	25.00	18.80	61.80
	Fear	38.46	10.70	58.04
	Joy	57.04	56.90	90.35
	Neutral	78.64	79.10	90.14
	Surprise	56.96	57.70	81.40
	Sad	28.12	36.90	83.55



	Angry	Disgust	Fearful	Joy	Neutral	Sad	Surprise
Angry	56	3	1	14	139	0	30
Disgust	9	35	0	0	9	15	0
Fearful	0	1	41	8	0	0	0
Joy	0	0	1	36	1	0	21
Neutral	1	0	5	1	23	5	1
Sad	24	0	21	1	1	23	1
Surprise	0	0	14	0	1	0	19

Figure 5. 5: Confusion matrix of the proposed Inception V3 DenseNet model on MELD dataset.

Figure 5.5 shows the confusion matrix of the proposed model used on the MELD dataset. The figure demonstrates that the proposed model displayed a strong performance in accurately labeling the input features. The proposed model demonstrated accurate recognition of the neutral emotion among the classes, with minimal times of misclassifications. The class defined by anger exhibits the highest rate of misclassification, and this emotional condition is frequently misquoted with neutral emotion. Furthermore, the feeling of sadness frequently gets mistaken for the emotions of anger and fear. Despite a certain incidence of misclassifications, the proposed model demonstrated higher accuracy compared to the other models on the MELD dataset.

Figure 5.6 compares the proposed model's precision to existing models. The figure shows that the proposed model is more precise than others. The proposed model is more precise for class joy than fear. MMFA has better precision than the other models. The precision metric shows the model's exact positives. The proposed model yields more true positives than the others. The proposed model has 79.22% average precision, higher than the other models. The model's discrimination helped achieve results.

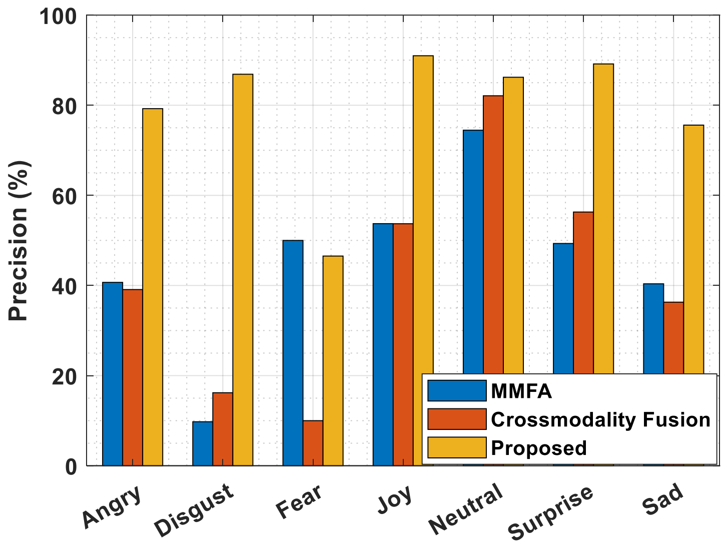


Figure 5. 6: Comparative analysis of precision for the proposed and existing models.

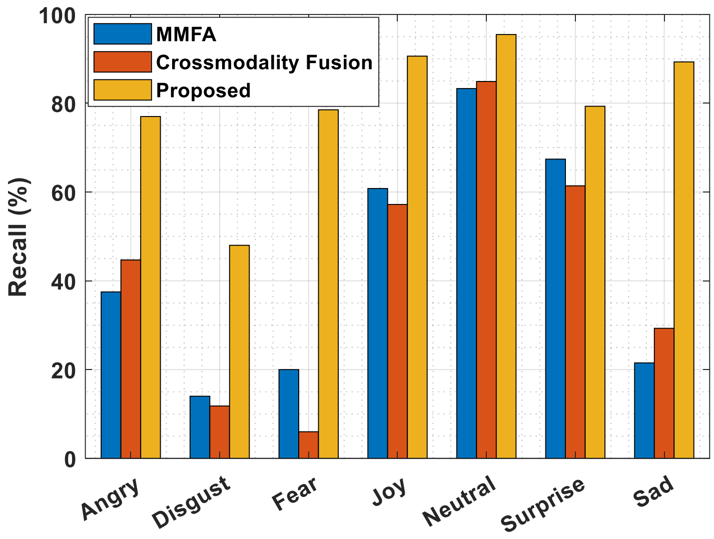


Figure 5. 7: Comparative analysis of recall for the proposed and existing models.

Figure 5.7 provides a comparative analysis of recall for the proposed and existing models. The depicted figure demonstrates that the proposed model attained superior recall values in comparison to the other models. Moreover, the proposed model has a higher recall rate for the neutral class and a lower reliability rate for the repulsive class. In general, the suggested model exhibits greater accuracy in precisely identifying the input features when compared to the current models. The average recall achieved by the proposed model has been determined to be 79.74%.

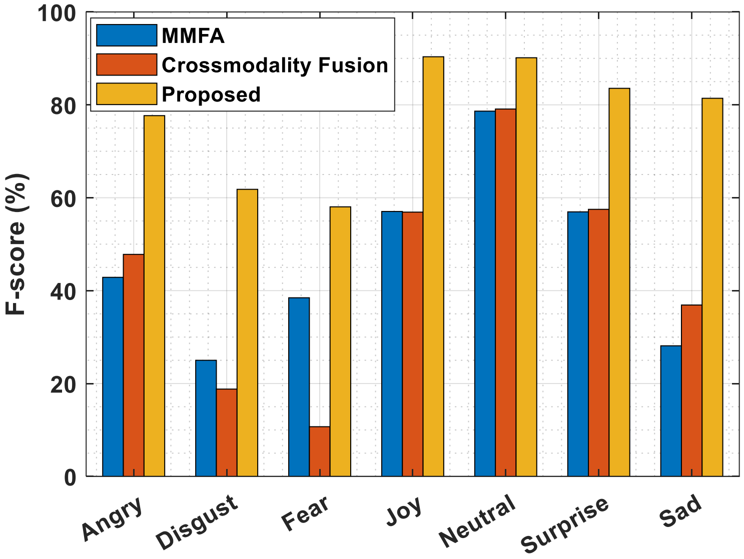
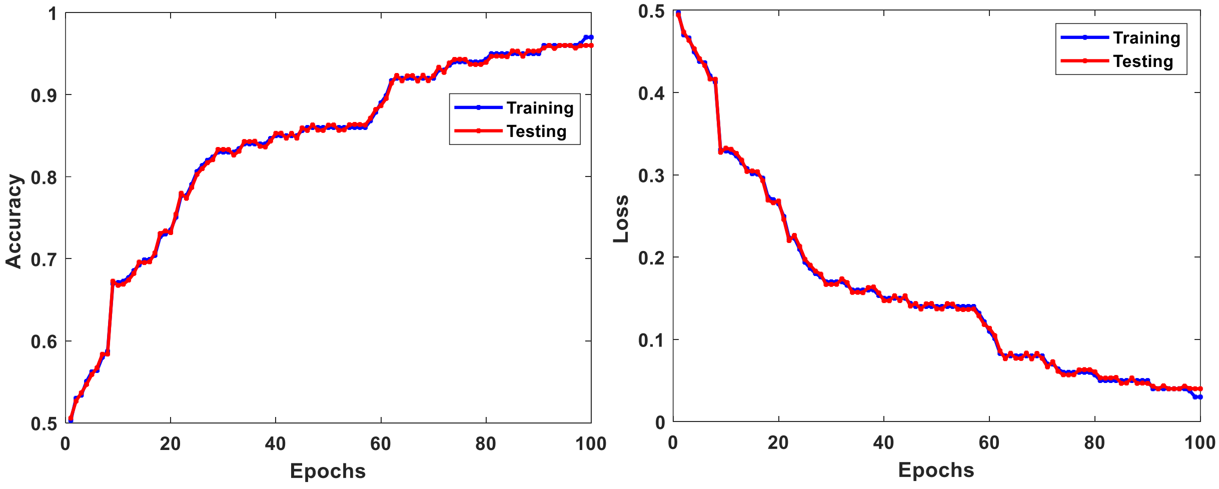


Figure 5. 8: Comparative analysis of F-score for the proposed and existing models.

The proposed and existing models' F-scores are compared in Figure 5.8. The F-score values for all classes show that the proposed model outperformed the existing models in Figure 5.8. Additionally, the model consistently scores well for joy and poorly for fear. F-score statistics showed that cross-modality fusion outperformed other models. The model outperformed others due to its selective ability. The proposed model has a 77.56% mean F-score.



(a)

(b)

Figure 5.9: (a) Model accuracy (b) model loss curves of the proposed model for training and testing data.

Figure 5.9 illustrates the accuracy and loss trends for the training and testing datasets. Figure 5.9(a) illustrates the model's accuracy, clearly indicating that the proposed model demonstrates better accuracy on both the training and testing datasets. The model keeps its fundamental level of accuracy up to 50 epochs, and at 100 epochs, it gets a lot better. Figure 5.9(b) also shows the model loss, which is a number that shows how much the model has lost over time. For both the training and testing datasets, the values have been plotted over 100 iterations. The initial stage of the model exhibits a significantly elevated loss. However, it steadily dropped to the lowest value as the number of epochs increased. Furthermore, the plot does not reveal any overfitting problems, and the model shows great accuracy in identifying the labels for the input features.

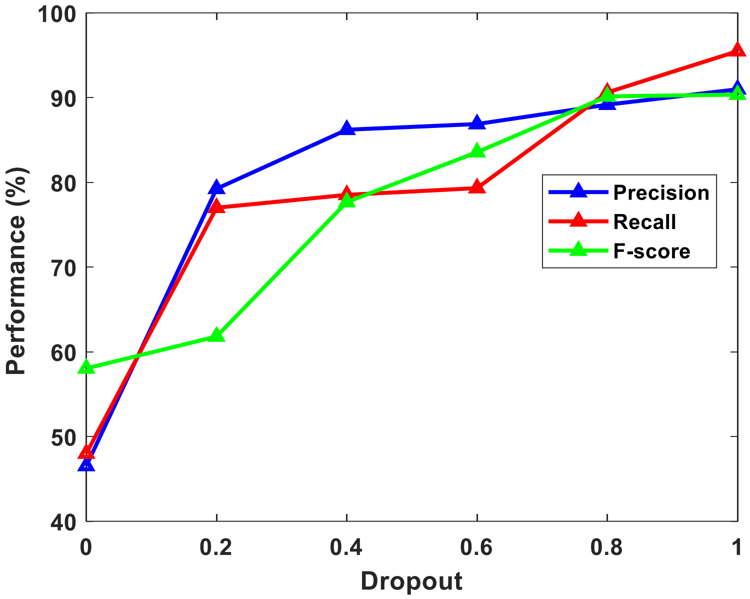


Figure 5. 10: Performance of the proposed model on different dropout rates.

Figure 5.10 shows a graph that illustrates the proposed model's performance as the dropout rates are varied. According to the provided figure, the selected dropout value for the proposed model is optimal since the model produces better outcomes when the dropout rate is set to 1. While the model made the lowest values without dropout, the performance metrics consistently increased as the dropout rate increased. Variations in the mini-batch size have been used to plot the accuracy of the model in Figure 5.11.

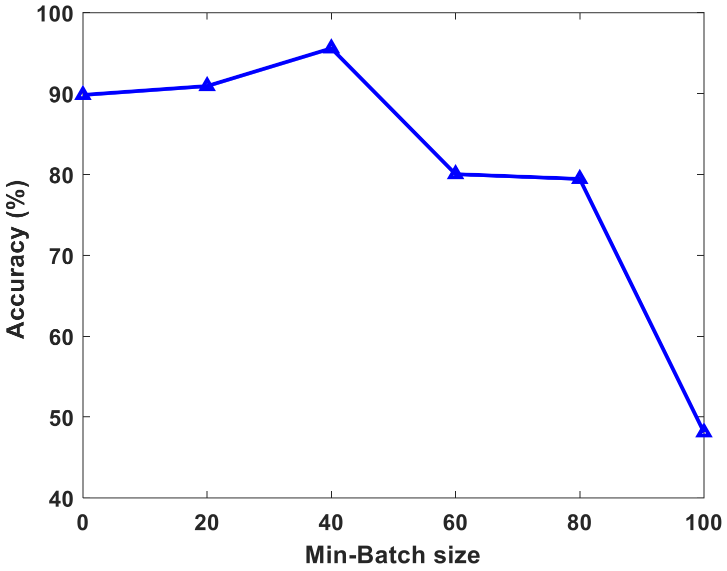


Figure 5. 11: Accuracy of the proposed model varying the mini-batch size.

The results shown in Figure 5.12 demonstrate that the proposed method achieves greater efficiency when using a mini-batch size of 40. Elevating the mini-batch size to 100 leads to a significant decline in accuracy. The optimal selection for training the proposed model on the MELD dataset has been determined to be a mini-batch size of 40. Upon reaching 40, the model's accuracy experiences a substantial decrease.

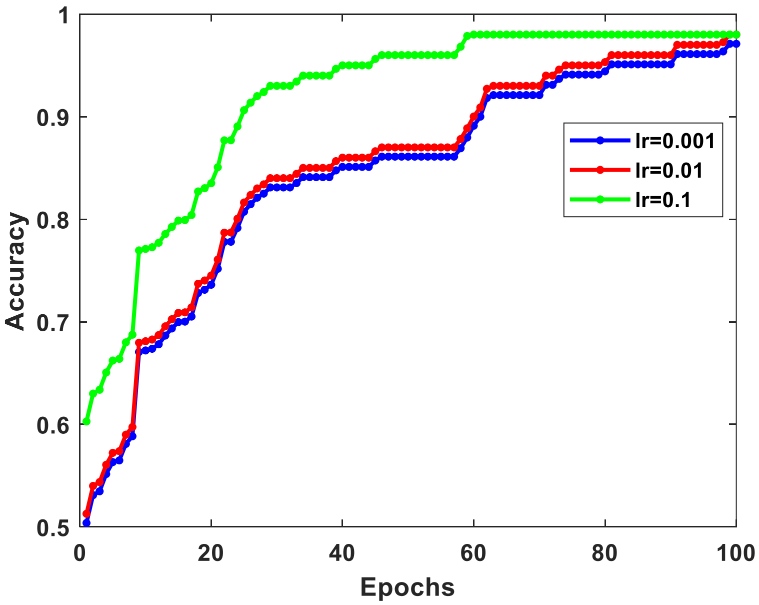


Figure 5. 12: Comparison of accuracy for the proposed model across various learning rates.

Figure 5.12 shows quantitative verification of the proposed model's accuracy for different learning rates. The figure shows that the proposed model was most accurate at 0.1 learning rate. The approach has been iterated with learning rates of 0.001 and 0.01. The model's accuracy decreased with different learning rates. When the learning rate is 0.01, accuracy drops significantly. After further reduction, the model had the lowest accuracy at various epochs. Thus, the model learning rate affects accuracy. In addition, the proposed model's emotion classification learning rate of 0.1 is much higher.

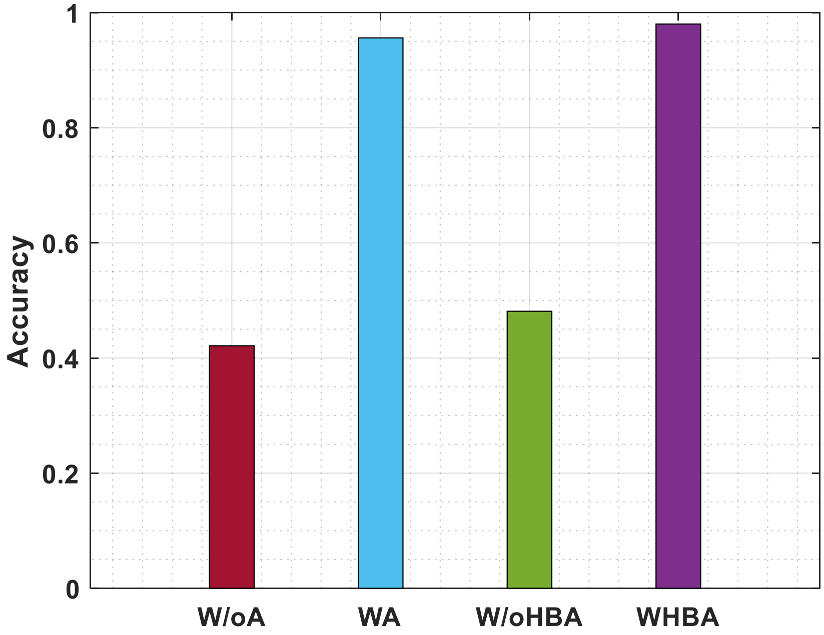
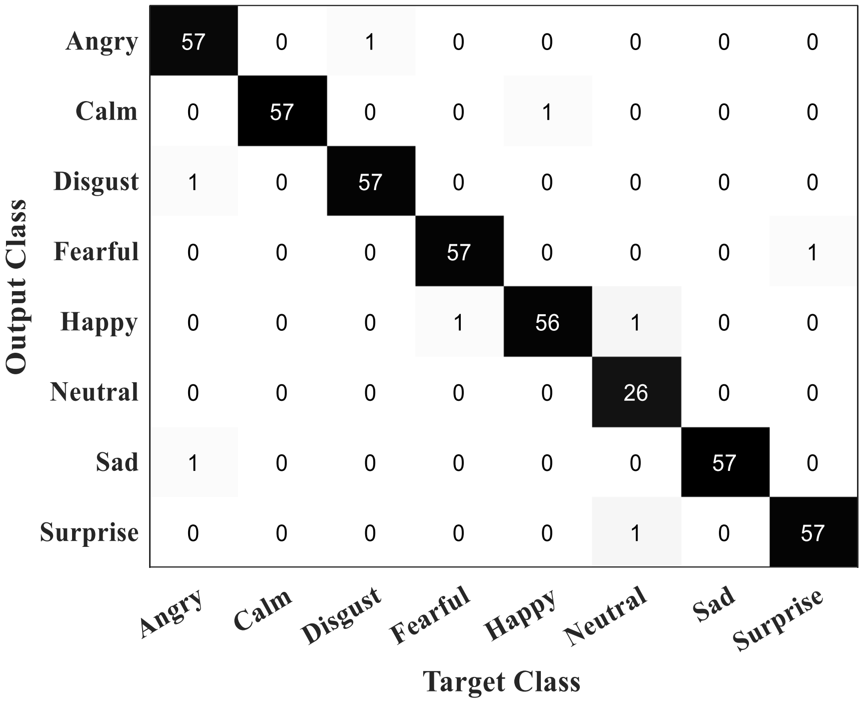


Figure 5. 13: Ablation experiments performed on the proposed model using the MELD dataset.

Ablation experiments have been conducted to demonstrate the significance of each module, and the results are shown in Figure 5.13. Figure shows the classification model with and without attention, labeled as W/oA and WA, respectively. WHBA and W/oHBA are classification models with and without HBA algorithms. The attention mechanism and weight tuning mechanism are crucial for achieving the desired results, as confirmed by the results.

b. RAVDESS dataset

The RAVDESS dataset is used to evaluate the proposed model, and the results are discussed below:



	Angry	Calm	Disgust	Fearful	Happy	Neutral	Sad	Surprise
Angry	57	0	1	0	0	0	0	0
Calm	0	57	0	0	1	0	0	0
Disgust	1	0	57	0	0	0	0	0
Fearful	0	0	0	57	0	0	0	1
Happy	0	0	0	1	56	1	0	0
Neutral	0	0	0	0	0	26	0	0
Sad	1	0	0	0	0	0	57	0
Surprise	0	0	0	0	0	1	0	57

Figure 5. 14: Confusion matrix for the proposed InceptionV3DenseNet model.

The proposed model's confusion matrix from the RAVDESS dataset is displayed in Figure 5.14. The figure makes it very clear that the model that was suggested is the best way to accurately identify emotions. With the proposed model, it is possible to accurately identify almost all classes. There is very little wrong classification in the RAVDESS dataset. The proposed model's ability to discriminate greatly reduced the rate of wrong classification and allowed accurate labeling of the features. The HBA algorithm also lowered the rate of classification errors by choosing the best weight values for predictions. On the whole, the proposed model was better at detecting emotions on the RAVDESS dataset.

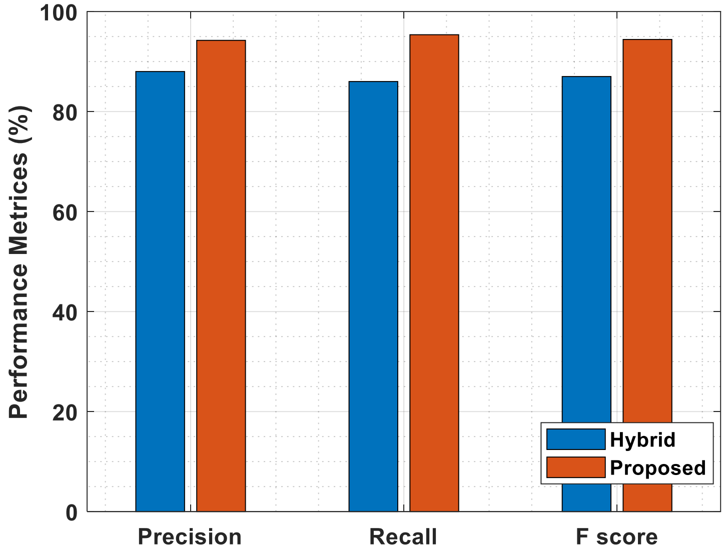


Figure 5. 15: Performance comparison of the proposed and existing models.

Figure 5.15 shows a comparison of the way the proposed model and other models work. The figure demonstrates that the proposed model achieved better values compared to the other models under comparison. Under its discriminative capacity, the model successfully distinguished the characteristics and delivered precise labels. Furthermore, the error rate is reduced using optimal training. It is observed that the proposed model achieves an average precision of 94.23%, an average recall of 95.36%, and an average F-score of 94.41%.

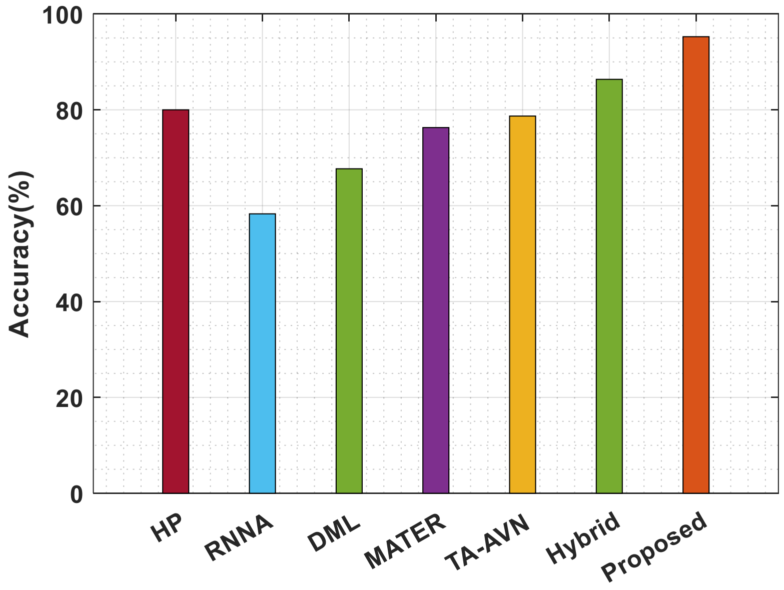
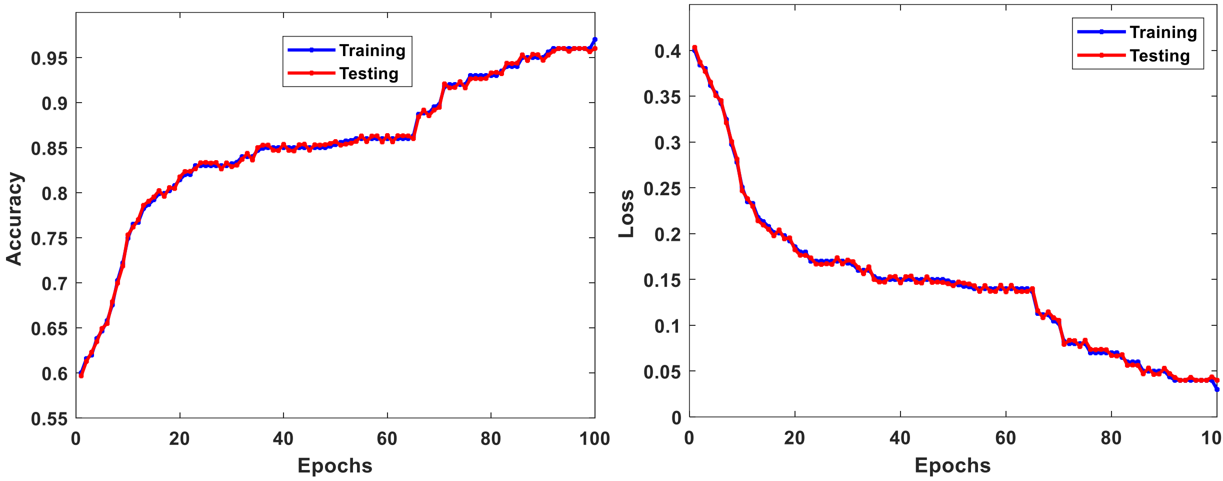


Figure 5. 16: Accuracy comparison of the proposed and existing models.

Figure 5.16 shows a comparison of the proposed model's accuracy with a number of other models that are already out there. The proposed model was the most accurate of the models that were already out there, as shown in the figure. The model's ability to tell the difference between the input features and the drop in error rate brought about by optimal tuning made it easier for the model to do what it was designed to do. The proposed model achieves an average accuracy of 95.25%.



(a)

(b)

Figure 5. 17: (a) Model accuracy and (b) model loss curves of the proposed model on training and testing data.

Figure 5.17 shows training and testing set model accuracy and loss estimates. Figure 5.17(a) shows 100-epoch estimated model accuracy for training and testing datasets. The graph shows the proposed model detects emotions better. The model performed better in training and testing datasets. The model reached peak accuracy after 100 iterations. Figure 5.17(b) shows training and testing dataset model loss. The analysis of the figure indicates that the model's loss is exceptionally low, registering a value below 0.05 after 100 iterations.

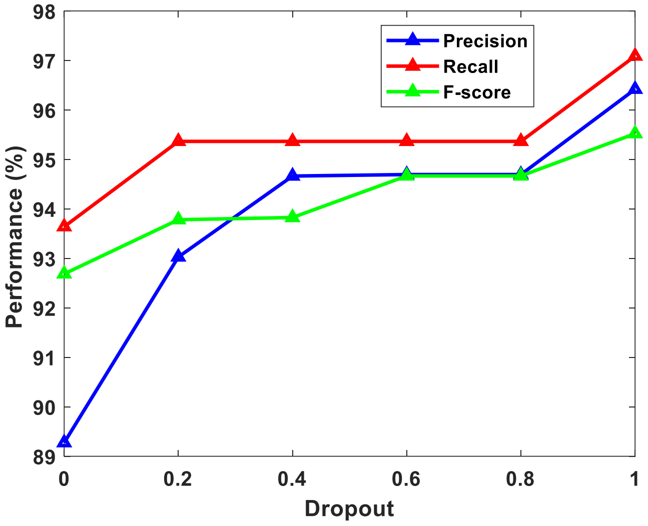


Figure 5. 18: Performance comparison of the proposed model on different dropout rates.

Figure 5.18 illustrates the efficacy of the proposed model across different dropout rates on the RAVDESS dataset. An analysis of the figure reveals that the performance values exhibit an upward trend as the dropout rate is raised from 0 to 1. Significant variations in the performance are observed when the dropout rate is adjusted in relation to the dropout rates.

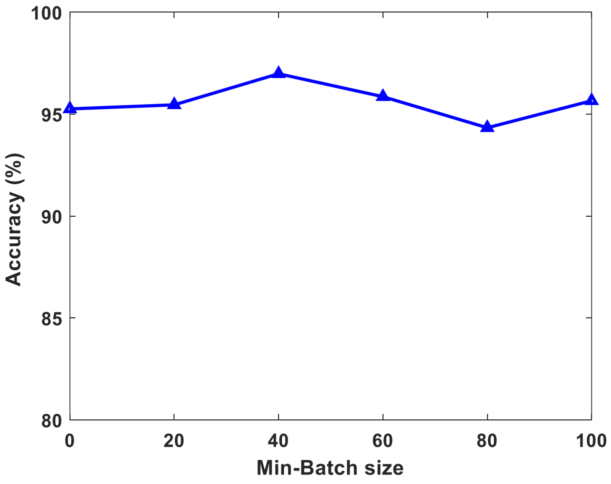


Figure 5. 19: Proposed model's accuracy across various mini-batch sizes.

The model's accuracy can be evaluated by varying the mini-batch size and displaying the plotted results in Figure 5.19. When the mini-batch size was set to 40, the proposed model achieved the highest level of accuracy, as shown in the figure. A noted reduction in accuracy occurs when the mini-batch size is adjusted to 40, whether increased or decreased.

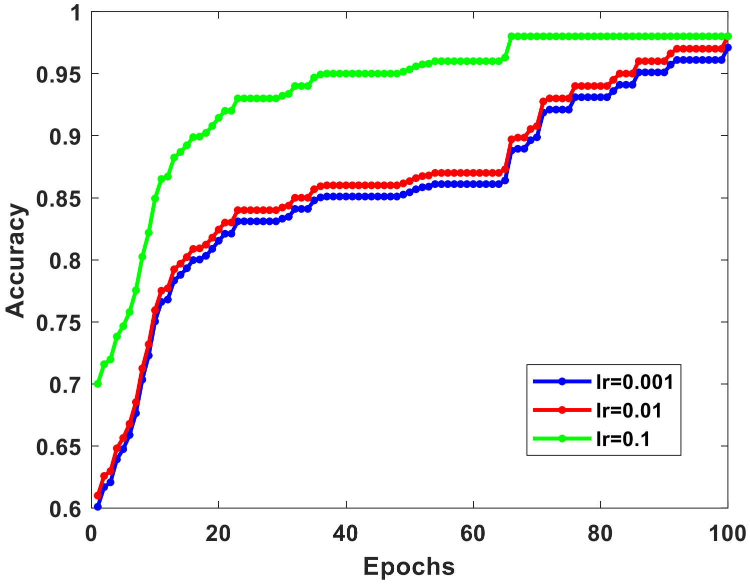


Figure 5. 20: Accuracy of the proposed model on different learning rates.

A plot showing accuracy has been generated for the model using various learning rates on the RAVDESS dataset. The analysis of Figure 5.20 indicates that the proposed model achieved optimal accuracy at a learning rate of 0.1. A notable reduction in accuracy occurs when the learning rate is lowered to 0.01, with a further decline evident at a learning rate of 0.001.

5.3.3 Ablation experiments

Ablation experiments are carried out to establish the proposed method's performance effectiveness and highlight the significance of various modules. In Figure 5.21, the accuracy of the model is shown to be low upon removing the attention and weight tuning mechanisms. By employing the tuning and attention mechanisms, the model achieves its highest level of accuracy.

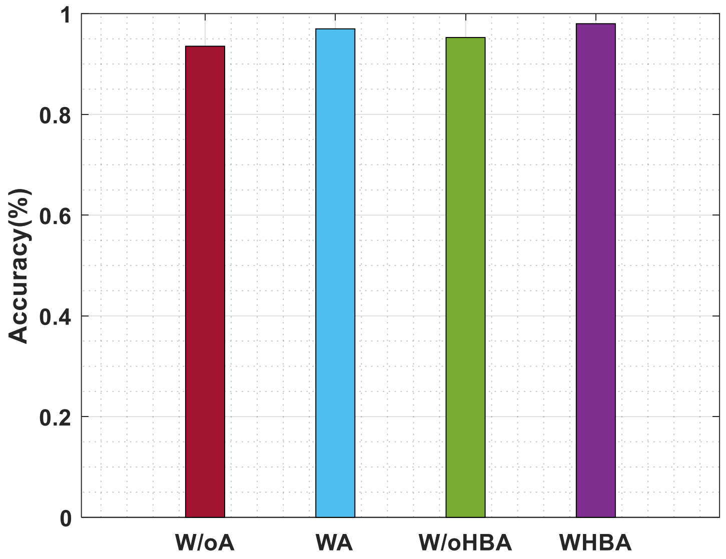


Figure 5. 21: Ablation experiments of the proposed model on the RAVDESS dataset

Simulations show that the proposed model can accurately identify the input features. The model's selective capacity enabled the achievement of the desired outcomes. Moreover, the attention mechanism integrated after the model improved its capacity for identifying emotions. The model's HBA algorithm optimized weight values and reduced classification error. Based on MELD and RAVDESS dataset simulations, the model outperforms most existing methods in accuracy.

Figure 5.22 shows a comparative analysis of the performance between the proposed and existing models. The figure demonstrates that the proposed model attained better performance metrics relative to the alternative model. The model's discriminative ability allowed it to proficiently differentiate features and assign precise labels. The error rate is reduced via optimal training. The proposed model attains an average precision of 94.23%, an average recall of 95.36%, and an average F-score of 94.41%.

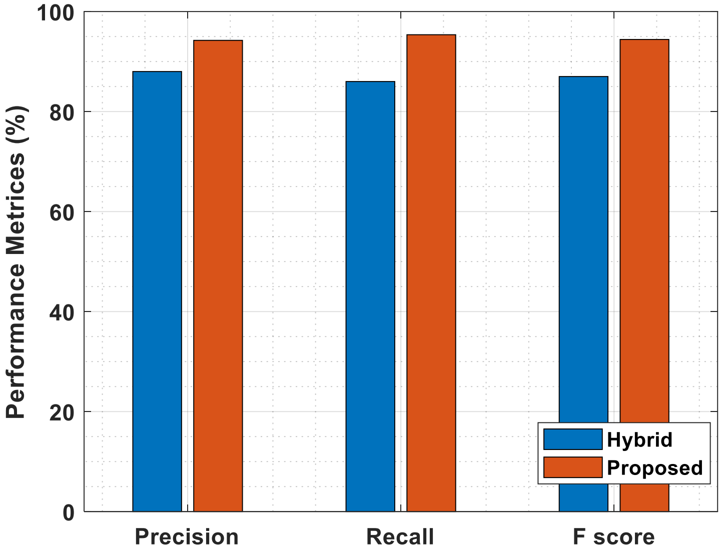


Figure 5. 22: Performance comparison of the proposed and existing models.

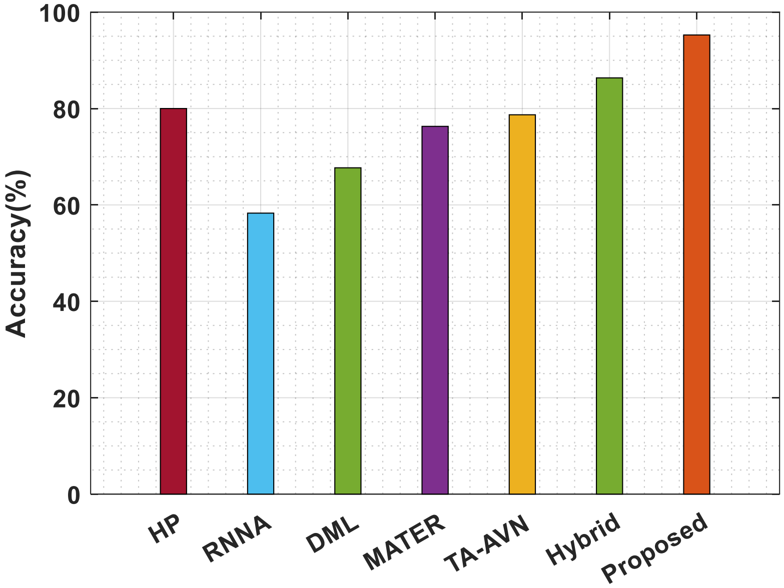


Figure 5. 23: Accuracy comparison of the proposed and existing models.

Figure 5.23 depicts the outcomes of the proposed model's accuracy evaluation in comparison to numerous existing models. The model was able to achieve the desired results by reducing the error rate through precise tuning and identifying input features. The average accuracy of the proposed model is 95.25%. The proposed method shows enhanced precision in comparison to existing methods. It surpasses the current methodologies, such as HP, RNNA, MATER, TA-AVN, and Hybrid approaches. It demonstrates the efficacy of the proposed model. The precision of current methodologies for emotion recognition typically ranges from 60% to 80%. It exceeds 90% for both the hybrid and proposed methodologies.

Moreover, the significance of each module in the suggested methodology is substantiated by the ablation experiments. The results were derived from multiple scenarios, and the model demonstrated higher accuracy relative to existing models in the literature across all considered cases. The research indicates that the proposed model achieves an overall accuracy of 74.87% on the MELD dataset and 95.25% on the RAVDESS dataset.

5.4 Summary

The proposed method can detect emotional cues from three distinct sources: audio, video, and text. Data is collected from the MELD and RAVDESS datasets, and specific attributes associated with each mode are identified. The features that have been discovered are used to predict labels. The study introduces a significant improvement by introducing a deep learning-based Inception V3DenseNet model for classification tasks. The proposed model exhibits higher accuracy than others, demonstrating its effectiveness in addressing classification challenges. The HBA algorithm is used to modify the model's weights, leading to a notable improvement in categorization accuracy by selecting optimal weight values. An attention mechanism is incorporated at the classifier's endpoint to highlight important features from various modalities. Integrating various elements into the model to analyze emotions through EEG signals has proven useful.

Chapter 6: ASDnet: Classification Model for Individuals with Autism Spectrum Disorder

6.1 Introduction

Treating and detecting ASD in its early stages is critical, and it helps decrease the symptoms to some extent and increase the quality of life for an individual. Early identification in childhood is essential because it can help children with ASD communicate and interact with others more effectively, improving their quality of life. Early disease identification is essential for illness management and treatment. Detecting autism at an early stage is a difficult task where several cases have occurred in the world. ASD could not be healed since its primary detection is necessary as it permits highly active mitigating treatment. Some techniques are used in automated facial expression recognition, and they help people with ASD become aware of their emotional states and react to them more efficiently. The recognition of facial expressions is challenging due to the wide range of expressions that can be expressed and the fact that facial emotions are natural. To solve this difficulty, most studies finalist and decrease the problem into six basic emotions: disgust, sadness, surprise, fear, anger and happiness [94].

The behavior of ASD is like an enigma in making meaningful movements, poor facial expressions, lack of name response, not saying simple words for 16 months, and more. The human face shows the key to facial emotion recognition. Even when affected with ASD, children show unpredictable facial emotions. As a result, it is difficult for families and counselors to categorize their facial expressions. Analyzing facial emotions solves ASD in the starting stage by using the kid's expressions through automated devices [95].

This chapter introduces the Dual-branch CNN-based visual transformation model (Db-CNN-VTM) to identify children with autism. In the beginning, the images from the dataset are pre-processed to get rid of noise. A geometric data augmentation (DA) is used to add more training data and pre-process. After pre-processing, feature extraction is performed by the Convolutional Neural Network with a grid-wise attention mechanism (CNN-GAM) model. After the feature extraction, the feature fusion residual network model is established. Finally, the classification is done using a novel Db-CNN-VTM model, and then the hyper parameter is found using the Hunger Game Search Optimization (HGSO) method.

This chapter aims to develop a new recognition model for detecting individuals with autism spectrum disorder. The contributions of this chapter are listed as follows:

a. To create and develop a real-time dataset of special children for the first time in Bangladesh.
b. To introduce a novel ASDnet classifier model based on facial Grid-wise expression features for recognizing children with autism spectrum disorder.
c. To learn the long-range dependencies between different facial regions by introducing a grid-wise attention mechanism in ASDnet.
d. To improve the speed of ASDnet by removing the impact of huge variations of scales in pyramid features using dual-branch transformation and fusion models.
e. To improve the accuracy of the proposed ASDnet by selecting optimal hyper parameters using metaheuristic approaches.
f. To validate the effectiveness of the proposed model by collecting real-time images from Autistic individuals.

6.2 Proposed Method

This chapter proposes a new Db-CNN-VTM model for classifying facial expressions into different groups. There are four steps in the proposed method: pre-processing, feature extraction, feature fusion, and classification. The first step is to choose images from the dataset and use the Geometric data augmentation (DA) method to prepare them. Then, the CNN-GAM (convolutional neural network with a grid-wise attention mechanism) is used to pull out the main geometric and appearance base features. After the features are extracted, they are sent to feature fusion, which combines the different features to get the most useful information. The feature fusion is carried out using the residual network. Finally, the dual-branch CNN-based visual transformation model (Db-CNN-VTM) performs the classification. This model will initially extract pyramid features from the high-level convolutional blocks. Then, these features will be split into two clusters, and each cluster will use a feature pyramid fusion (FPF) method to obtain a single-scale feature map from multi-scale features. After obtaining the feature map from the FPF module, a visual transformation-based attention (ViTA) mechanism is integrated with deep level convolutional filters for enhancing the long-range bias learning. Finally, this Db-CNN-VTM uses a normalized, fully connected network to recognize children with ASD. The hyper parameters of the proposed deep learning models are optimized using the Hunger Games Search optimization (HGSO) algorithm. The proposed method workflow is explained in Figure 6.1.

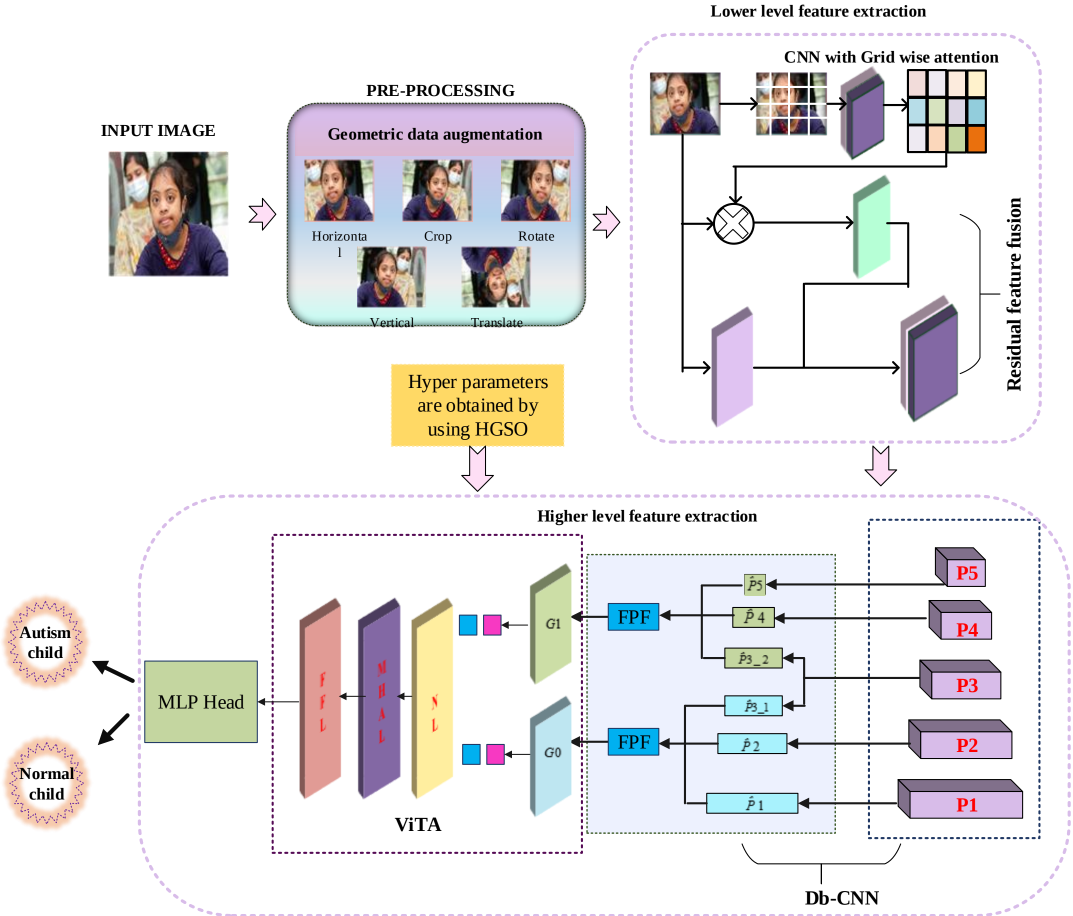


Figure 6. 1: Capturing the images of special children for the ASDnet dataset.

6.3 Dataset Description

Dataset 1: Dataset 1 is a publicly available Special Children's Facial Image Data Set at https://www.kaggle.com/general/123978. Here, Images of children with autism were collected from websites dedicated to the disorder, while images of typical children were randomly gathered from the Internet. The training dataset includes 1327 images of children who had autism and 1327 images of children without autism. The test set contains 140 facial images of children diagnosed with autism and 140 images of children not confirmed with autism.

Dataset 2*:* The real-time images utilized in this paper were collected from children aged between 4 and 13. It includes 283 images of ASD children and 331 images of normal children. The images of children with ASD are captured from the Smiling Children Special School, Dhaka, Bangladesh. Normal children were selected from regular schools (not special education schools). Experts in behavioral evaluation have already pre-diagnosed ASD in kids. Children with other neurological conditions and those on medication were not allowed to participate in the study. Before the study, each child's parent gave their informed consent.

Two special schools have agreed upon legal agreements to use images. The images of children with ASD are captured from the special schools below:

a. Proyash Institute of Special Education, Dhaka Cantonment, Dhaka-1200.
b. Smiling Children Special School, Aftabnagar, Kamal House, House No# 40, Road- 06, Block- E, Sector - 01, Dhaka 1212.

The images were captured using a Canon M50 camera with a 24.1-megapixel APS-C CMOS sensor and an EF-M mount. The Canon EF-M Mount 15-45mm f/3.5-6.3 zoom lens is a standard component of the M50 kit. All the photographs were taken in portrait layout to represent the children's facial expressions precisely. The images vary in size from 5 to 10 megabytes, which is deemed satisfactory for a camera with a resolution of 24.1 megapixels. The dataset is shared in the github platform for future researchers. Available Link: https://github.com/mashukalamgir/Autism

Figure 6. 2: Capturing the images of special children for the ASDnet dataset.

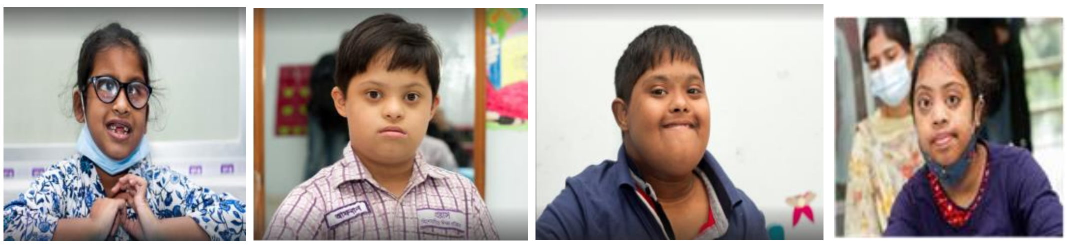


Figure 6. 3: Sample images from Dataset 2.

6.4 CNN with a Grid-wise Attention Mechanism for Local Feature Extraction

The CNN-GAM generates a weighted feature map from low-level convolutional blocks. In general, the aligned facial images are referred to as $R^{\tilde{C}\times\tilde{H}\times\tilde{W}}$, where $\tilde{C}$ is the image channel), $\tilde{H}$ is the size, and $\tilde{W}$ is the depth. After cropping the image, the image grids of size $\tilde{h}\times\tilde{w}$ is forwarded to the feature extraction module. The grid selection is represented as:

$$GRID\left(R,\tilde{h},\tilde{w}\right)=\left\{R_{1,1}^{\tilde{C}\times\frac{\tilde{H}}{\tilde{h}}\times\frac{\tilde{W}}{\tilde{w}}},......,R_{i,j}^{\tilde{C}\times\frac{\tilde{H}}{\tilde{h}}\times\frac{\tilde{W}}{\tilde{w}}},.......,R_{\tilde{h},\tilde{w}}^{\tilde{C}\times\frac{\tilde{H}}{\tilde{h}}\times\frac{\tilde{W}}{\tilde{w}}}\right\}=R^{\tilde{h}\tilde{w}\times\tilde{c}\times\frac{\tilde{H}}{\tilde{h}}\times\frac{\tilde{W}}{\tilde{w}}} \tag{6.1}$$

The input image $\boldsymbol{R}$ is split into $\tilde{h}\times\tilde{w}$ grids, $R_{i,j}^{\tilde{C}\times\frac{\tilde{H}}{\tilde{h}}\times\frac{\tilde{W}}{\tilde{w}}}$ represents the grid image with the shape of $\tilde{C}\times\frac{\tilde{H}}{\tilde{h}}\times\frac{\tilde{W}}{\tilde{w}}$, and situates in the i^{th} row and the j^{th} column in the image grids of R $\left(1\le i\le\tilde{h},1\le j\le\tilde{w},and\,i,j\in N\right)$. To study the local features of the facial region in the grid, each $R_{i,j}$ will be sent to a local network for feature extraction. The low-level feature extraction network (LLFEN) is represented as,

$$\hat{R}_{i,j}=Low\left(R_{i,j}\right),\hat{R}^{h_i w_i\times C_i\times\frac{H_i}{h_i}\times\frac{W_i}{w_i}}=Low\left(R^{h_i w_i\times C_i\times\frac{H_i}{h_i}\times\frac{W_i}{w_i}}\right) \tag{6.2}$$

Where $\hat{R}_{i,j}$ denotes the feature map attained using LLFEN. These feature maps are given as input to a grid-wise attention mechanism for assigning the weights to these feature maps based on their importance.

6.4.1 Grid-wise Attention

This block uses a matrix dot product unit to determine the similarity between the grids. The following formulation describes this procedure:

$$\begin{gathered}\Re_k=\frac{\tilde{W}}{\tilde{w}},Q=\hat{R}^{\tilde{h}\tilde{w}\times\tilde{C}\times\frac{\tilde{H}}{\tilde{h}}\times\frac{\tilde{W}}{\tilde{w}}},key=\hat{R}^{\tilde{h}\tilde{w}\tilde{C}\times\frac{\tilde{W}}{\tilde{w}}\times\frac{\tilde{H}}{\tilde{h}}},scores=\frac{Q*key}{\Re_k},\\ attention=softmax\left(scores\right)=\hat{R}^{\tilde{h}\tilde{w}\times\tilde{C}\times\frac{\tilde{H}}{\tilde{h}}\times\frac{\tilde{W}}{\tilde{w}}}\end{gathered} \tag{6.3}$$

Here, $*$ denotes the matrix dot product. Here, adaptive pooling squeezes every channel into a scalar form and expands it back to its real size. The process can be formulated as follows:

$$APool_{avg} = adaptiveavgpool2d((1,1)),\ \hat{R}^{\tilde{h}\tilde{w}\times\tilde{C}\times 1\times 1} = APool_{avg}(attention),\ expansion = Ones\left(\frac{\tilde{H}}{\tilde{h}},\frac{\tilde{W}}{\tilde{w}}\right),$$
$$pattern = expansion \otimes \hat{R}^{\tilde{h}\tilde{w}\times\tilde{C}\times 1\times 1} = \tilde{R}^{\tilde{h}\tilde{w}\times\tilde{C}\times\frac{\tilde{W}}{\tilde{w}}\times\frac{\tilde{H}}{\tilde{h}}} \tag{6.4}$$

Where $\otimes$ depicts the scalar matrix multiplication, $adaptiveavgpool2d((1,1))$ is used to convert a matrix operand into a scalar form. $Ones\left(\frac{\tilde{H}}{\tilde{h}},\frac{\tilde{W}}{\tilde{w}}\right)$ is utilized for generating a matrix of size $\left(\frac{\tilde{H}}{\tilde{h}},\frac{\tilde{W}}{\tilde{w}}\right)$ with all ones. An attention matrix is generated by concatenating the weights of each grid pattern to create a global attention matrix.

$$\tilde{R}^{\tilde{C}\times\tilde{H}\times\tilde{W}} = Ungrid\left(\tilde{R}^{\tilde{h}\tilde{w}\times\tilde{C}\times\frac{\tilde{H}}{\tilde{h}}\times\frac{\tilde{W}}{\tilde{w}}}\right),\ \tilde{R}'^{\tilde{C}\times\tilde{H}\times\tilde{W}} = \tilde{R}^{\tilde{C}\times\tilde{H}\times\tilde{W}} = R^{\tilde{C}\times\tilde{H}\times\tilde{W}} \tag{6.5}$$

Where $Ungrid$ is a process that inverts these grid attentions into the unique face image from $\boldsymbol{R}$. Thus, $\tilde{R}'^{C_i\times H_i\times W_i}$ at the low-level feature learning stage, long-range distortions between different facial regions are considered while constructing a feature map.

6.5 Pre-processing

Pre-processing is done by geometric DA methods (including translation, random rotation, cropping, horizontal reflection, and vertical reflection) by adding additional training data to the proposed model that can be generalized. The pre-processed images are shown in Figure 6.3.

a. *Translation:* The original image is moved randomly along the horizontal or vertical axis. The sides of the images are padded with zeros.

b. *Random rotation:* Random rotation is applied to the image between -90° and $+90^{\circ}$.

c. *Cropping:* The cropping technique randomly samples a section of an image. The cropped image shows a relevant part of the face.

d. *Horizontal reflection:* Also known as horizontal flip, a horizontal reflection is a transformation that reflects a figure across the y-axis.

e. *Vertical reflection:* This technique is called vertical flipping and creates a horizontal mirror image of the original image. Performing a vertical reflection is the same as rotating an image 180° and performing a horizontal reflection.

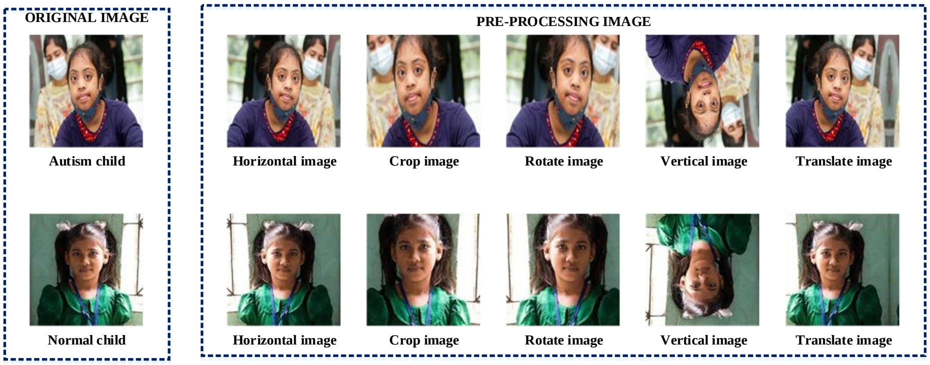


Figure 6. 4: Pre-processing of images.

6.6 Feature Fusion

The proposed model uses a residual network (RN) for combining the feature map $\tilde{R}'$ with the original image R. The original image $\boldsymbol{R}$ and the weighted feature map $\tilde{R}'$ are forwarded to their respective feature transformation networks (FTN), but their learning parameters are not shared. The transformed features are forwarded to the fusion network after the feature transformation. This process can be formulated as follows:

$$\overline{R}^{\tilde{C}\times\tilde{H}\times\overline{W}} = RN\left(FTN1(R)+FTN2\left(\tilde{R}\right)\right) \tag{6.6}$$

Where, $FTN1$ denotes the F.T.N. of the input image, $FTN2$ denotes the FTN of the weighted feature map $\left(\tilde{R}\right)$, and RN denotes the residual network. In this section, low-level features of facial images were extracted using a grid-wise attention mechanism. The obtained feature map $\tilde{R}^{\tilde{C}\times\tilde{H}\times\tilde{W}}$ is forwarded to the Db-CNN-VTM for high-level feature extraction and classification.

6.7 Db-CNN-VTM for High-level Feature Extraction and Classification

To classify individuals with autism spectrum disorder, the Db-CNN-VTM is proposed. Initially, this model extracts feature characteristics of the pyramid from high-level convolutional blocks. To obtain a multi-scale feature map from single features, these features are split into two clusters, each using Feature Pyramid Fusion (FPF). The feature map obtained by these FPF modules is used for generating visual tokens. Here, the ViTA mechanism is introduced for enhancing long-term bias learning. Finally, this Db-CNN-VTM uses the SoftMax layer to recognize children with ASD. In the proposed model, the hyper parameters are optimized using HGSO.

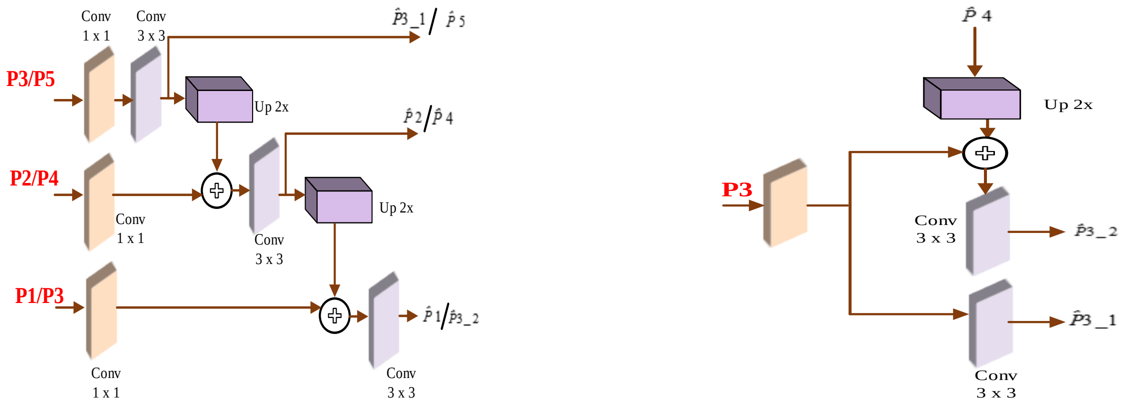
The proposed Db-CNN-VTM extracts pyramidal feature maps $P1$ to $P5$ using CNN-based pyramid network. Here, the features extracted by the backbone model are denoted as $P1$, $P2$ and $P3$. After that, 2 down-sample 3 × 3 convolution layers are used to generate $P4$ and $P5$ as shown in Figure 6.5. Here, high-level semantic features are extracted using the subsequent expression:

$$\hat{P}_i = \varphi\left(ups\left(\hat{P}_{i+1} \oplus \chi(P_i)\right)\right) \tag{6.7}$$

where $\hat{P}_i$ and $\hat{P}_{i+1}$ denote the present and deeper feature map correspondingly. χ denotes a 1×1 convolution layer and is used for reducing the channel of the input feature map P_i. ups denotes the process of a size of 2. Also, the component-wise addition is denoted as $\oplus$, and φ denotes a 3 × 3 convolution, and it is utilized to reduce an aliasing result of up-sampling. Here, the feature maps $\hat{P}5$, $\hat{P}4$ and $\hat{P}3\_2$ are merged to provide a new feature map $G1$ in the first branch. In the same way, the feature maps $\hat{P}3\_1$, $\hat{P}2$ and $\hat{P}1$ are merged to provide a new feature map $G0$ in the second branch. The following expressions are used to compute $\hat{P}5$ and $\hat{P}3\_1$:

$$\hat{P}_i = \varphi(\delta(P_i)) \tag{6.8}$$

Next, the feature maps $\hat{P}1, \hat{P}2$ and $\hat{P}3\_1$ are assigned to the small branch, $\hat{P}3\_2$, $\hat{P}4$ and $\hat{P}5$ to the large branch, as illustrated in Figure 6.5. Every branch uses a feature pyramid fusion (FPF) module for the generation of $G0$ and $G1$ using multiple-level features fusion.



(a) (b)

Figure 6. 5: Network structure (a) Pyramid feature extraction module (b) Structure for the generation of $\tilde{P}3\_1$ and $\tilde{P}3\_2$.

6.7.1 Feature Pyramid Fusion

The proposed model divides the feature pyramid into two sets and is allocated to two branches. Each branch fuses multi-level features into a single-scale feature map using FPF. The explanation of the FPF module is given by considering a tiny branch. The input for the FPF module is given as $\hat{P}1$ $\hat{P}2$ and $\hat{P}_{3\_1}$ and the output feature map $G0$. Initially, every input is rescaled into a similar size (the stride is 4). Element-by-element addition operations merge individual multi-level features with different respective fields. The feature map incorporates multi-level features after the fusion. However, this crude merging fails to smoothly combine the multi-scale data from many layers into a feature map. Applying a context module after the add operation can reduce the up-sampling aliasing effects and increase the field of interest, enhancing the fusion effect. The FPF module can merge multiple features with multi-scale data to create a single-scale feature map, as shown in Figure 6.6.

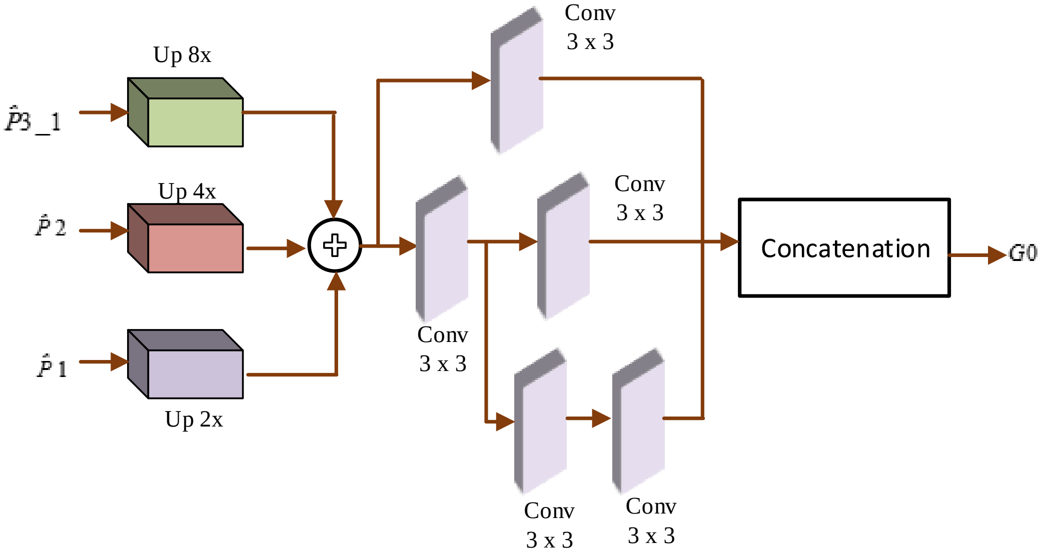


Figure 6. 6: Structure of the FPF unit.

6.7.2 Visual Transformation-based Attention (ViTA) Mechanism

The feature maps ($G0$ and $G1$) generated by the FPF module are converted into visual semantic tokens for fitting the input need of a visual transformer. Then, the feature maps are converted to take on the same form as $\tilde{C}_1 \times \tilde{H}_1 \times \tilde{W}_1$:

$$G_i'^{1\times\tilde{C}_1\times\tilde{H}_1\times\tilde{W}_1} = Re\,shape\left(G_i^{1\times\tilde{C}_1\times\tilde{H}_1\times\tilde{W}_1}\right), \quad i = 1,2 \tag{6.9}$$

Where G_i' denotes the visual token of respective feature maps G_i . These tokens can be embedded as:

$$\xi_i^{1\times D} = TEN\left(G_i^{1\times\tilde{C}_1\tilde{H}_1\tilde{W}_1}\right), \;\; i = 1,2 \tag{6.10}$$

where TEN denotes the Token embedded Network. The suggested method introduces a trainable embedding classification token to the input sequence, like the one used in Vision Transformer (ViT). The visual unit of the facial image is represented as:

$$\left(\xi_{cls}, \xi_1, \cdots \xi_i \cdots \xi_N\right), \quad i = 1,2,\cdots N \tag{6.11}$$

Where ξ_i denotes the visual token of size D. Here, a visual token embedding with a learnable 1-dimensional position is introduced to keep the positional information on pyramid features. The encoder of a typical Transformer receives the succeeding sequence of visual tokens as input. After the positional embedding, the visual token sequence can be described as follows:

$$A_0 = (\xi_{cls}; \xi_1, \cdots \xi_i \cdots \xi_N) + Embed_{position} \quad , i = 1, 2, \cdots N \tag{6.12}$$

The encoder that primarily consists of repeated blocks of multi-head attention (MHA) and multilayer perceptron (MLP) is then applied in a manner like that of ViT. The procedure is written as follows:

$$a_i' = MHA(a_{i-1}) + a_{i-1}, \quad i = 1,2,\cdots \gamma \tag{6.13}$$

$$a_i = MLP(a_i') + a_i', \quad i = 1,2,\cdots \gamma \tag{6.14}$$

where γ denotes the number of M.H.A. and M.L.P. repeated blocks. It is noted that each block's fully connected networks use a layer-normed method to prevent an over fitting issue. Finally, a layer-normed, fully connected network uses the global vector a_γ^0 to identify children with ASD.

$$y = ASD\left(NormLayer\left(a_\gamma^o\right)\right) \tag{6.15}$$

where k denotes the number of classes.

6.8 Model Training using Hunger Games Search Optimization

CNN-GAM and Db-CNN-VTM provide the learnable parameters for ASDnet. In this chapter, the loss function includes label smoothing to lessen over fitting as given as follows:

$$\ell = -\frac{\sum_{i=1}^{\beta} log\left(soft\left(y_i^{predicted}\right)\right) \times smoothing\left(y_i^{actual}\right)}{\beta} \tag{6.16}$$

where β denotes the batch size of the feed forward model. log and $smoothing$ denote the logarithmic function and a label smoothing function, respectively. Specifically, the $smoothing$ is described as:

$$smoothing(y_{hot}) = y_{hot}(1-\varepsilon)+\frac{\varepsilon}{k} \tag{6.17}$$

where y_{hot} denotes the one-hot encoding and ε denotes the constant parameter. The proposed model utilizes Hunger game search optimization (HGSO) to minimize the objective function. HGSO was inspired by the cooperative conduct of social animals, who search more actively when hungry. In this work, the position of individuals $X(t)$ in HGSO is related to the weight parameters of the proposed Db-CNN-VTM model. Initially, the weights of Db-CNN-VTM are initialized randomly. After that, the fitness of all the populations is computed using (6.16). The best fit, worst fit, and location of the best individual in the current iteration are updated after sorting the solutions based on the fitness value. The position of the individual is then iteratively updated using the subsequent expression.

$$\overrightarrow{X(T+1)} = \begin{cases} \overrightarrow{X(T)}.(1+randn(1)), & r1<l \\ \overrightarrow{W_1}.\overrightarrow{x_z}+\vec{R}.\overrightarrow{w_2}.\left|\overrightarrow{X_z}-\overrightarrow{X(t)}\right|, & r1>l,r2>E \\ \overrightarrow{W_1}.\overrightarrow{x_z}-\vec{R}.\overrightarrow{w_2}.\left|\overrightarrow{X_z}-\overrightarrow{X(t)}\right|, & r1>l,r2<E \end{cases} \tag{6.18}$$

Where, $\overrightarrow{X(t)}$ indicates the location of individuals, $\overrightarrow{X_z}$ indicates the position of the best individual, $\overrightarrow{W_1}$ and $\overrightarrow{W_2}$ are the weight of the individual, $\vec{R}$ value is among $[-a,a]$, $r1$and $r2$ are uneven numbers among [0,1], $randn(1)$ indicate a random numbers casual distribution, T is iterations in current. l denotes the organized variable, it controls the algorithm sensitivity. E indicate differentiation control for all positions.

$$E = \sec h(|f(i)-best\cos t\ function|) \tag{6.19}$$

Where, $f(i)$ denote the cost of each population function, $i \in 1,2,....n$, and $\sec h$ denotes the hyperbolic function and is equal to $\left(\sec h(X)=\frac{2}{e^{x}+e^{-x}}\right)$. The $\vec{R}$ is expressed as,

$$\vec{R} = 2\times a\times rand - a \tag{6.20}$$

$$a = 2\times\left(1-\frac{t}{Max_{iter}}\right) \tag{6.21}$$

A random number between [0, 1] is denoted by $rand$, and Max_{iter} is the greatest number of iterations. The expression for $\overrightarrow{W_1}$ in equation (18) is defined as:

$$\overrightarrow{W_1(i)} = \begin{cases} hun_{gry}(i).\frac{M}{SHun_{gry}}\times r_4, & r3 < l \\ 1 & r3 > l \end{cases} \tag{6.22}$$

The expression $\overrightarrow{W_2}$ is described as follows:

$$\overrightarrow{W_2(i)} = (1-\exp(-\left|hun_{gry}(i) - SHun_{gry}\right|))\times r_5 \times 2 \tag{6.23}$$

The population size is indicated by M, the hunger of each population denoted as hun_{gry}, the amount of starving feelings of every population in $SHun_{gry}$ is denoted as $sum\left(hun_{gry}\right)$, and r_3, r_4 and r_5 denotes the uneven numbers between $[0,1]$. For each population, starvation is accurately modeled as:

$$hun_{gry}(i) = \begin{cases} 0, & All_{fitness}(i) == Best\cos t\ function \\ hun_{gry}(i)+T & All_{fitness}(i) != Best\cos t\ function \end{cases} \tag{6.24}$$

Where $AllFitness(i)$ reserves the fitness of everyone in the present iteration. T is represented by the expression.

$$T = \begin{cases} pt\times(1+r), & ht < lh \\ ht, & ht \geq lh \end{cases} \tag{6.25}$$

$$ht = \frac{f(i)-best}{wf-best}\times r_6 \times 2\times(U_{\lim it} - L_{\lim it}) \tag{6.26}$$

where, T is limited to lesser bound pt, r represents an uneven number connecting [0,1], wf and $best$ represents the best and worst fitness achieved all through the iteration present, respectively, $f(i)$ represents the fitness of every individual population, r_6 is an uneven number between [0,1], $U_{\lim it}$ and $L_{\lim it}$ are upper and lower limits of the dimension, respectively. This procedure is repeated to a certain number of iterations to get the optimal weights. Figure 6.7 explains the flow chart of the HGSO algorithm.

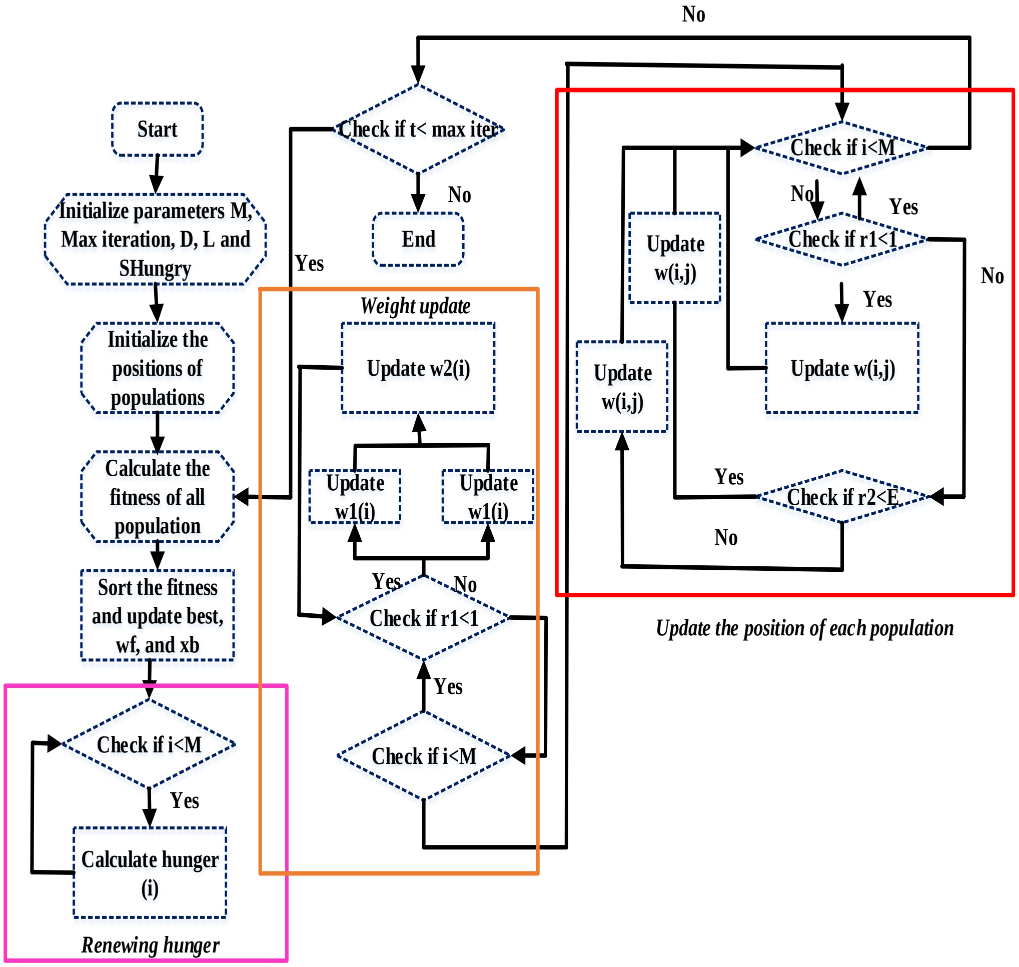


Figure 6. 7: Flow chart of HGSO algorithm.

6.9 Experimentation and Result Analysis

The proposed ASDnet for recognizing children with autism is simulated using the Python programming language. In this section, the performance of the proposed model is validated by comparing it with other models qualitatively and quantitatively on publicly available and collected ASD datasets. The proposed ASDnet model uses CNN-GAM for local and global feature extraction. Also, a new Db-CNN-VTM is introduced for final classification. Here, dual branch CNN acts as a backbone network to generate visual semantic tokens. ViTA receives the sequence obtained from this backbone network. ViTA has a depth of 12. The heads of MHA and the dimension of the project key vector are considered as 8 and 64, respectively. The proposed ASDnet uses 100 epochs and a batch size of two for its training. The suggested ASDnet model's weight parameters are optimized using the HGSO. Additionally, a 0.0005 drop in the learning rate is considered. An ablation study is conducted to show the efficiency of each part of our model.

6.9.1 Performance Metrics

Precision, specificity, recall, accuracy, f-measure, and Kappa coefficient are the main performance measures that are maximized in the proposed method. Here, classification accuracy is measured to find the best model for differentiating autism-related facial emotions from normal children. The total number of images in the dataset is denoted as n . t_p, t_n, f_p, f_n are represented by the true positive, true negative, false positive and false negative. Table 6. 1 represents the performance metrics formulation. The computations for the performance metrics under consideration are as follows:

6.9.2 Performance Evaluation

The proposed ASDnet model uses a Db-CNN-VTM classifier to classify the input images as an autistic child and a normal child based on facial expressions. Figure 6.8 shows the proposed ASDnet's real-time dataset confusion matrix.

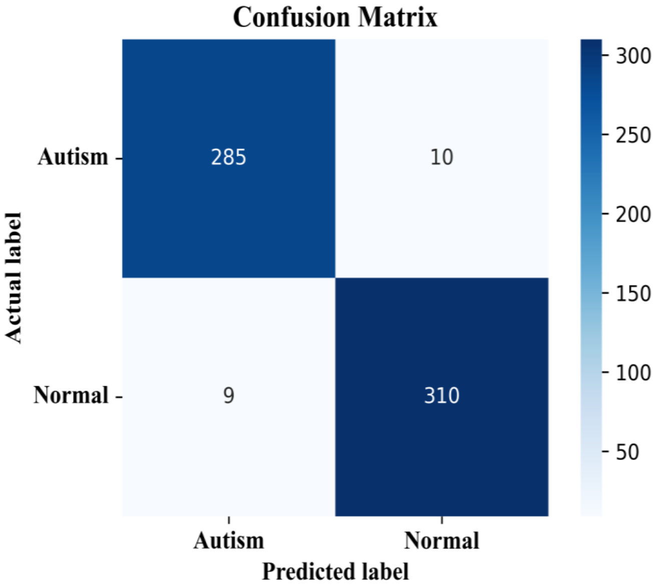
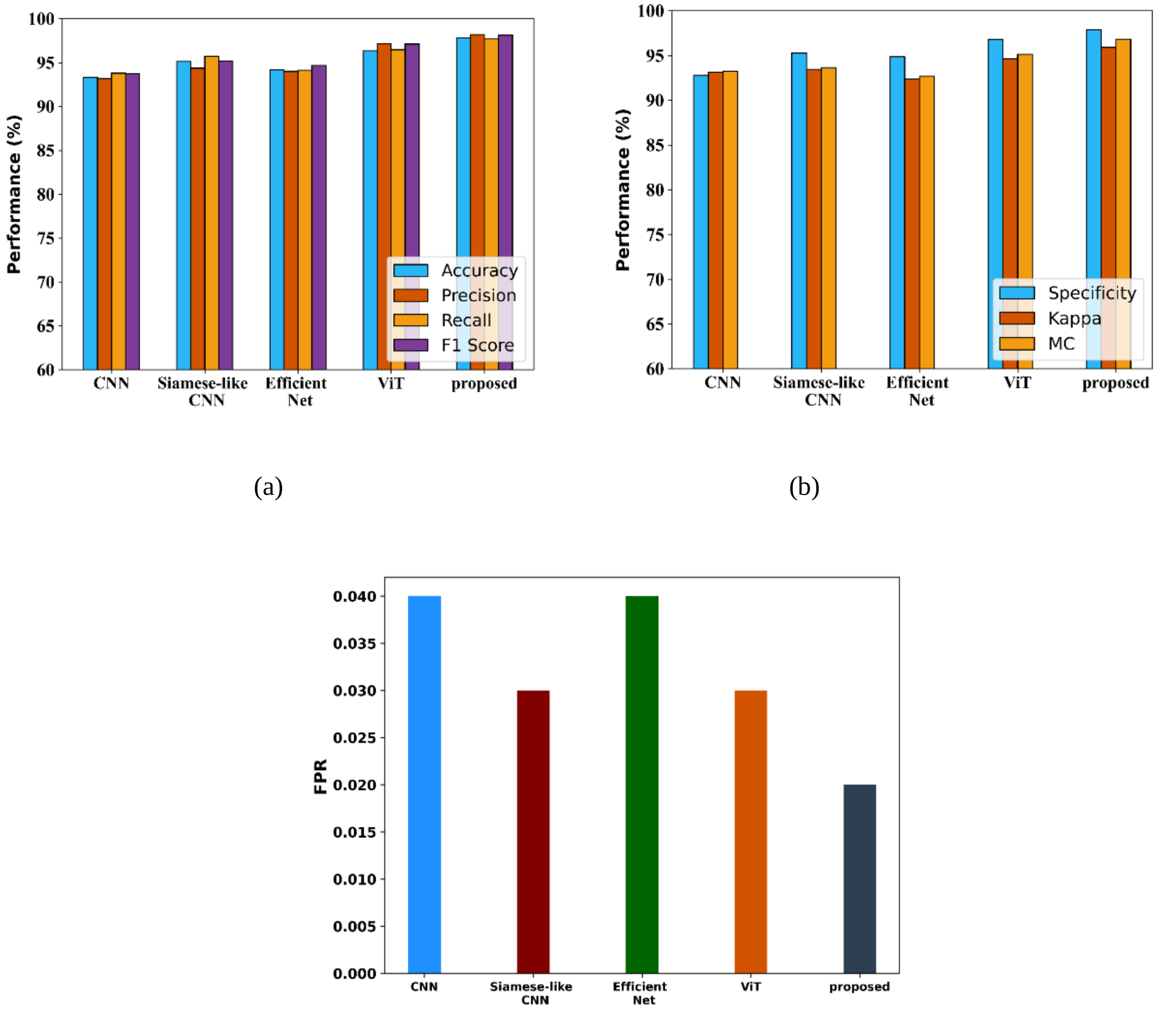


Figure 6. 8: Confusion matrix for Dataset 2 (real-time).

Here, the performance of the proposed model is validated by comparing it with other baseline models such as CNN (Tang, 2018), Siamese-like CNN (Chen, 2021), ViT (Cao, 2023) and EfficientNet (Mujeeb Rahman, 2022). In this analysis, the most popular existing models are all implemented to compare the same datasets fairly. The effectiveness of the proposed model is compared with other networks on Dataset 1 and Dataset 2 in Tables 6.2 and 6.3, respectively. Figures 6.8 and 6.9 compare the proposed model's performance to existing models.

Table 6. 1: Comparative analysis on Dataset 1.

Methods	CNN (Tang, 2018)	Siamese-like CNN (Chen, 2021)	EfficientNet (Mujeeb Rahman, 2022)	ViT (Cao, 2023)	**Proposed**
Accuracy	93.34	95.14	94.17	96.34	**97.82**
Precision	93.16	94.36	93.99	97.15	**98.18**
Recall	93.79	95.72	94.13	96.47	**97.73**
F1-score	93.73	95.18	94.68	97.12	**98.12**
Specificity	92.78	95.27	94.89	96.78	**97.89**
FPR	0.04	0.03	0.04	0.03	**0.02**
Kappa	93.14	93.42	92.36	94.62	**95.92**
MC	93.24	93.64	92.67	95.14	**96.83**



(a)

(b)

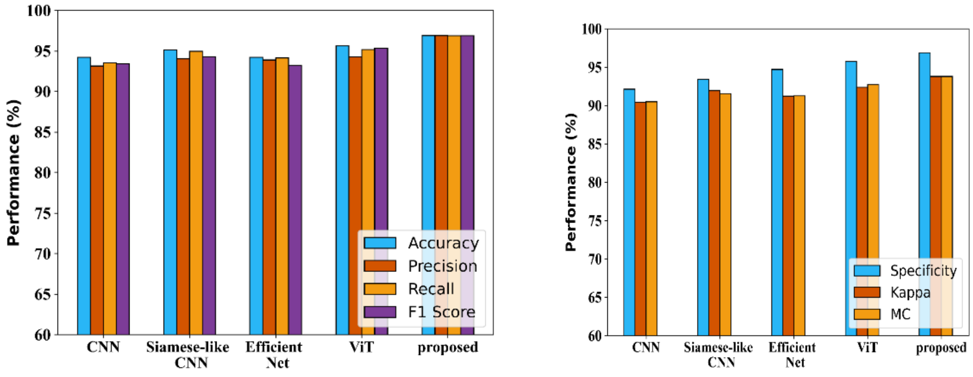
(c)

Figure 6. 9: Comparative analysis on Dataset 1.

(a) Accuracy, Precision, recall, and F1-score (b) Specificity, Kappa, and MC (c) F.P.R.

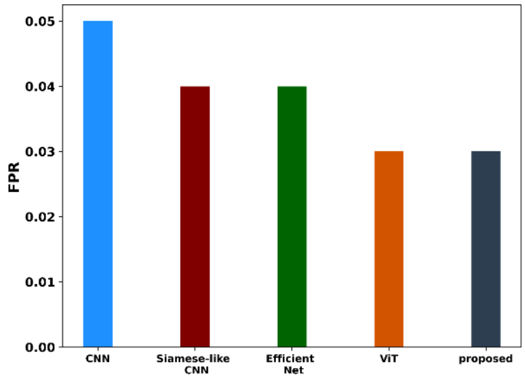
Table 6. 2: Comparative analysis on Dataset 2.

Methods	**CNN (Tang, 2018)**	**Siamese-like CNN (Chen, 2021)**	**EfficientNet (Mujeeb Rahman, 2022)**	**ViT (Cao, 2023)**	**Proposed**
Accuracy	94.23	95.14	94.23	95.63	**96.91**
Precision	93.14	94.06	93.89	94.27	**96.91**
Recall	93.53	94.97	94.16	95.18	**96.89**
F1-score	93.42	94.28	93.23	95.34	**96.9**
Specificity	92.14	93.42	94.71	95.76	**96.89**
FPR	0.05	0.04	0.04	0.03	**0.03**
Kappa	90.42	91.95	91.23	92.36	**93.8**
MC	90.49	91.56	91.27	92.71	**93.8**



(a)

(b)



(c)

Figure 6. 10: Comparative analysis on Dataset 2

(a) Accuracy, Precision, recall, and F1-score (b) Specificity, Kappa, and MC (c) FPR.

Tables 6.1 and 6.2 shows that the proposed ASDnet outperforms all other models on the two datasets. It provides accuracy of 97.82% and 96.91 on Dataset 1 and Dataset 2, respectively. Here, the classification accuracy of CNN (Tang, 2018) and EfficientNet (Mujeeb Rahman, 2022) models is reduced due to the lack of learning long-range inductive biases between various facial regions. Also, when head posture deviation or partial occlusion happens, the Siamese-like CNN (Chen, 2021) cannot reasonably predict emotion category and intensity. As a result, its performance is reduced on the ASD dataset since it holds an image of a child with Head pose deviation**.** According to this analysis, the ViT (Cao, 2023) performs better than all other models because it extracts data from the entire image. However, ViT needs a large-scale dataset for learning spatial inductive biases. However, the collected ASD dataset has a limited number of images. Alternatively, the proposed model uses dual CNN to learn long-range biases in the high-level semantic feature. Hence, it achieved high accuracy even for smaller ASD datasets.

Selecting the best weight parameter for improving the classification's accuracy is a frequent problem with all existing deep-learning models. As a result, the existing models display poor classification accuracy. As an alternative, the proposed model is combined with the HGSO method for optimal weight parameters. The effectiveness of the HGSO in the ASDNet model is confirmed by plotting accuracy and loss plots, as shown in Figure 6.11. The proposed ASDNet model with HGSO needs only 18 epochs for convergence. It shows the efficiency of HGSO in model training.

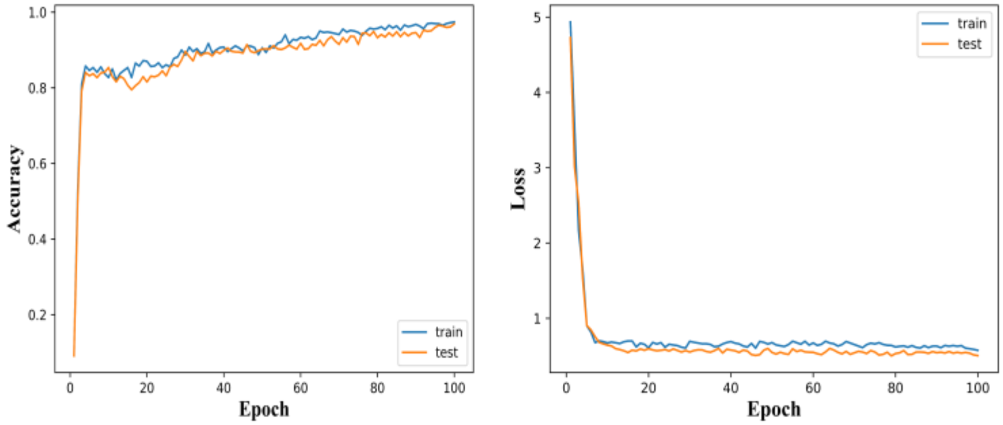


Figure 6. 11: Accuracy and loss curve of ASDnet dataset.

The proposed model is also validated using Receiver Operating Characteristic (ROC) curves. The ROC curve is selected based on the true positive rates (TPR) and false positive rates (FPR). The ROC curve graph is explained in Figure 6.12. The TPR increases gradually as the FPR is changed. For the real-time dataset, the proposed ASDnet model achieves an area under the curve (AUC) of 96.8% for ASD recognition.

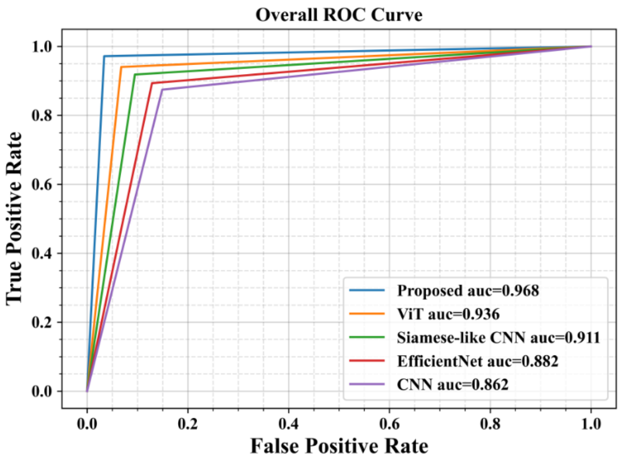


Figure 6. 12: ROC curve for ASDnet dataset.

6.9.3 Ablation Study

This section conducts a thorough ablation study to show how well the different parts of the suggested ASDnet model work. Table 6.4 examines the contributions of different ASDnet components using real-time datasets. CNN+GAM surpasses the backbone CNN by achieving a precision of 94.36 %, recall of 93.94%, and accuracy of 94.67%. Here, convolutional filters are guided by the grid-wise attention mechanism to learn richer characteristics of the facial contour profile and textures for ASD recognition. The CNN+GAM+ ViTA model outperforms the combination of the CNN+GAM model. Here, the convolutional filters in the deeper layer can concentrate on the crucial facial units and remove potentially distracting noise with the aid of the ViTA mechanism.

On the other hand, Dual-branch CNN performs better than baseline (single-branch CNN). This finding proves that dual branches are more effective than single branches. There are faces with smaller than 10 pixels and larger than 1000 pixels in the same dataset, and the baseline (single branch CNN) doesn't perform well in this task. Dual-branch CNN improves the detection of multi-scale faces while reducing scale changes in A.S.D. recognition. It is important to note that the accuracy of the combination GAM+Db-CNN+ ViTA has significantly increased to 96.43% on the real-time dataset. In the combination GAM+Db-CNN+ ViTA +HGSO, HGSO gives an additional level of confidence by selecting the weights optimally. It confirms combining HGSO and GAM+Db-CNN+ ViTA is effective for the ASD classification.

Table 6. 3: Ablation study on the ASDnet model components.

Model components	**Precision**	**Recall**	**Accuracy**
CNN	93.14	93.53	94.23
CNN+GAM	94.36	93.94	94.67
CNN+GAM+ ViTA	95.67	95.16	95.21
GAM+Db-CNN+ ViTA	96.12	96.03	96.43
GAM+Db-CNN+ **ViTA +HGSO**	**96.91**	**96.89**	**96.91**

6.10 Summary

This chapter proposed a novel ASDnet classifier model to recognize ASD based on a child's facial emotions. The proposed model introduced grid-wise attention for low-level feature learning and visual transformation-based attention for high-level feature learning. These attention mechanisms have been utilized to extract the relationships between different regions from facial images, and they also use a series of visual semantic tokens to learn the global representation. Also, the backbone CNN network has been replaced with Dual branch CNN (Db-CNN) to obtain multi-scale features. Also, it improves the convergence speed of the training process by integrating the network model with HGSO. The simulation results on two datasets (i.e., publicly available and real-time) show that the proposed model outperforms the state-of-the-art performances both quantitatively and qualitatively. The performance measurements confirmed the performance analysis of the proposed ASDnet by attaining an accuracy of 96.91%, precision of 96.91%, recall of 96.89%, F1-score of 96.90%, and specificity of 96.89% on the real-time dataset. The present study is limited to certain conditions: when hair, glasses, and caps partially obscure the face, occlusions may provide significant difficulty in recognizing emotions.

Chapter 7: Conclusion and Future Scope

Given the increasing preference for artificial intelligence in developed countries for detecting or predicting complex mental health problems, it's crucial to urgently address the disparity in South Asian countries, including our motherland. These countries are not yet equipped to meet the demands of emerging AI-based applications for mental health care. The current mental health care system, which is predominantly man-powered and manual, has its limitations, as discussed in the introductory chapter. This dissertation advocates a few alternative architectures to aid the manual process of making any sophisticated decision for identifying mental health problems such as ASD.

7.1 Concluding Interpretations

This thesis proposes several novel methods for automatically identifying facial emotions based on images from various datasets. In Chapter 3, the suggested study offered a unique hybrid DBRO approach for more accurate categorization. The selected features in which classification is performed and randomly initialized weight parameters are improved by an optimization algorithm. The method was evaluated on seven emotions, and experimental results showed that the proposed technique efficiently classified input feature images into the correct emotion category. The total performance of the method has presented an accuracy of 97% value that was not reported in any other classification model.

In Chapter 4, we proposed and implemented a revolutionary Bi-ENN to get the best classification results, which is used to categorize people's emotions in the FER system more precisely. The proposed system has the ability of extracting emotional signals from three distinct modalities: audio, video, and text. The first step involves collecting input data from the MELD and RAVDESS datasets and extracting pertinent features related with the modalities. The retrieved features are subsequently integrated using the MICCA algorithm to enhance feature discrimination. HOG and GF are used to extract significant characteristics from images, such as geometric and appearance-based features. After feature extraction, the dimensionality of the features is decreased by feature selection, with the MOSOA method, which was proposed in Chapter 3, used to choose the best features.

The simulation results showed that the proposed Neural Network surpassed other existing classification methods regarding accuracy. The proposed methodology accurately assessed the emotions, as the context provided for categorization was enhanced through training. The proposed Bi-ENN approach is especially appropriate for disease classification in the medical domain, as well as for FER systems. The Bi-ENN framework is extremely flexible and can effectively address any classification challenge with minimal alterations to the training process. The proposed Bi-ENN will be implemented in real-time for various categorization challenges in the future. The Bi-ENN is sufficiently adaptable to handle any classification challenge with minimal modifications to the training process. Consequently, the goal is to conduct multiple studies for classifying individuals' facial expressions using the Bi-ENN for real-time data analysis. The classifier uses the integrated features for label prediction. The proposed study makes an important contribution by presenting a deep learning-based InceptionV3DenseNet model for classifications. When compared to the other models, the suggested model had the greatest accuracy value.

In Chapter 5, we proposed a novel structure capable of extracting emotional cues from three modalities: audio, video, and text. The input data is retrieved from the MELD and RAVDESS datasets, and the modalities of these features are extracted. The extracted features are subsequently integrated using the MICCA algorithm to enhance the distinctions among the features. The merged features are subsequently input into the classifier to produce a class label for the facial input. This work's primary contribution is the application of InceptionV3DenseNet for classifications using a deep learning model. It achieved the highest accuracy among all proposed models. The model weight optimization was conducted using the HBA algorithm. This step enhanced accuracy by selecting the appropriate weight value in classification. The constructed model was simulated under different conditions on the MATLAB platform and evaluated using two standard benchmark datasets. The results from the simulation experiments show the proposed model can accurately label the input features due to its better discrimination capability. The model achieves an overall accuracy of 74.87% and 95.25% over the MELD and RAVDESS datasets, respectively.

Table 7. 1: Comparative analysis of All classification models.

DATASET	Methods	Precision (%)	Recall (%)	F-measure (%)	Accuracy (%)	Specificity (%)	Kappa (%)	FPR (%)	MCC (%)	MSE (%)
JAFFE	DBRO	98.07	98.57	98.25	98.55	98.96	94.08	0.24	98.31	1.45
KDEF	DBRO	98.20	98.78	98.46	98.41	98.96	92.74	0.24	98.49	1.59
JAFFE	Bi-ENN	98.57	98.70	98.57	98.57	99.77	93.66	0.32	97.65	1.33
CK+	Bi-ENN	98.90	98.57	98.68	98.75	99.79	91.53	0.26	96.35	1.22
MELD	Inception V3-Densenet	79.22	79.74	77.56	74.87	91.26	x	0.01	x	x
RAVDESS	Inception V3-Densenet	94.23	95.36	94.41	95.25	96.36	x	0.01	x	x
Dataset-1 (Normal Child)	Db-CNN-VTM	96.91	96.89	96.90	96.91	96.89	93.80	0.03	93.80	0.03
Dataset-2 ASDnet (Special Child)	Db-CNN-VTM	96.93	96.89	96.90	96.93	96.89	93.80	0.03	93.80	0.03

In Chapter 6, we present a new facial expression database specifically for children with Autism Spectrum Disorder (ASD), which captures a range of emotional expressions. It was the first-ever database for special children, focusing on Bangladesh only. Since there are similarities in facial features of all the countries of the South Asia region, this dataset can be utilized for detecting ASD in this region mentioned. We also outlined a Dual-branch CNN-based model, Db-CNN-VTM, for identifying Autism Spectrum Disorder through facial emotion recognition, achieving an impressive accuracy of 96.91%. It employs a Transformer-based Convolutional Neural Network (Tr-CNN) for effective feature extraction and classification of facial expressions in autistic children, surpassing existing models like ResNet and DNN. We also highlighted the advancements in automated facial expression recognition and the challenges faced by individuals with autism in expressing emotions.

To sum up, FER in deep learning is a very hopeful frontier to identify mental health problems by analyzing emotional expressions. With CNNs and other advanced machine learning techniques, FER systems can identify subtle emotional changes related to conditions such as depression, anxiety, and other affective disorders. The study introduces a Dual-branch CNN-based model, Db-CNN-VTM, for identifying Autism Spectrum Disorder through facial emotion recognition, achieving an impressive accuracy of 96.91%. It employs a Transformer-based Convolutional Neural Network (Tr-CNN) for effective feature extraction and classification of facial expressions in autistic children, surpassing existing models like ResNet and DNN. The research highlights advancements in automated facial expression recognition and the challenges faced by individuals with autism in expressing emotions.

7.2 Remaining Challenges and Future Work

The proposed methods in this dissertation addressed the major challenges involved in bringing the theory of deep learning-based classifier model design to practice and integrating it into the identification of mental health problems such as ASD. Below, we enumerate a few challenging issues that the dissertation does not solve. These are not fundamental limitations; however, we believe they can be solved within the improvised classification frameworks.

A. The proposed approaches cannot overcome the issue of occlusions, which are partially obscured by hair, glasses, and caps and may significantly complicate recognizing emotions. Occlusion-aware ASD recognition should be progressed by exploiting differential local features and efficient gated methods.

B. We could not implement the proposed ASD classification model using a multimodal approach, such as considering the audio, video, or text of special children and then predicting their classes appropriately. In the future, developers must address one modality, such as the visual features of a special child, for predicting his appropriate classes with high accuracy.

C. The proposed classification model will benefit our country's underprivileged people who have limitations in the early detection of mental health problems (ASD) due to some social taboos and financial issues. In the future, a mobile app-based solution or tool must be developed to overcome the social stigma, hence helping parents with early detection and undergoing appropriate therapies as early as possible.

D. The scalability needs to be improved, since deep learning models are data hungry. In the future, more datasets will be utilized to test the proposed model and accurately examine the computing problems. Aside from that, we plan to construct more optimum models that can improve algorithm training's global stability even after longer iterations with different data samples.

E. By combining real-time data and metadata, this system might be expanded in the future for emotion categorization in ASD patients. More development will be necessary for the app when the model's usefulness is tested in a clinical environment with children with Autism.

With the continued advance of research and technology, it is possible that soon FER will be able to play a major role in early detection and individualized interventions, therefore improving mental healthcare and quality of life for many people specifically the underprivileged ones in our country.

References

[1] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, "Comparing Recognition Performance and Robustness of Multimodal Deep Learning Models for Multimodal Emotion Recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 2, pp. 715–729, Jun. 2022, doi: https://doi.org/10.1109/tcds.2021.3071170.

[2] S. Gupta, P. Kumar, and R. K. Tekchandani, "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models," *Multimedia Tools and Applications*, Sep. 2022, doi: https://doi.org/10.1007/s11042-022-13558-9.

[3] A. J. Hong, D. DiStefano, and S. Dua, "Can CNNs Accurately Classify Human Emotions? A Deep-Learning Facial Expression Recognition Study," *arXiv (Cornell University)*, Jan. 2023, doi: https://doi.org/10.48550/arxiv.2310.09473.

[4] M. Ianculescu, A. Alexandru, and E.-A. Paraschiv, "The Potential of the Remote Monitoring Digital Solutions to Sustain the Mental and Emotional Health of the Elderly during and Post COVID-19 Crisis in Romania," *Healthcare*, vol. 11, no. 4, p. 608, Feb. 2023, doi: https://doi.org/10.3390/healthcare11040608.

[5] L. Garekwe, S. J. Ferreira-Schenk, and Zandri Dickason-Koekemoer, "Modelling Factors Influencing Bank Customers' Readiness for Artificial Intelligent Banking Products," *International journal of economics and financial issues*, vol. 14, no. 1, pp. 73–84, Jan. 2024, doi: https://doi.org/10.32479/ijefi.15238.

[6] E. Loth *et al.*, "Identification and validation of biomarkers for autism spectrum disorders," *Nature Reviews Drug Discovery*, vol. 15, no. 1, pp. 70–70, Dec. 2015, doi: https://doi.org/10.1038/nrd.2015.7.

[7] K. A. Shaw *et al.*, "Early Identification of Autism Spectrum Disorder Among Children Aged 4 Years — Early Autism and Developmental Disabilities Monitoring Network, Six Sites, United States, 2016," *MMWR. Surveillance Summaries*, vol. 69, no. 3, pp. 1–11, Mar. 2020, doi: https://doi.org/10.15585/mmwr.ss6903a1.

[8] M. O. Adebiyi, D. Fatinikun-Olaniyan, A. A. Adebiyi, and Abiodun Afolabi Okunola, "Survey on Current Trend in Emotion Recognition Techniques Using Deep Learning," *IEEE*, Apr. 2023, doi: https://doi.org/10.1109/seb-sdg57117.2023.10124548.

[9] S. Cooper, C. W. Hobson, and S. H. van Goozen, "Facial emotion recognition in children with externalising behaviours: A systematic review," *Clinical Child Psychology and Psychiatry*, vol. 25, no. 4, pp. 1068–1085, Jul. 2020, doi: https://doi.org/10.1177/1359104520945390.

[10] J. Shukla, "Empowering Cognitive Stimulation Therapy (CST) with Socially Assistive Robotics (SAR) and Emotion Recognition," *Tdx.cat*, 2018, doi: http://hdl.handle.net/10803/586279.

[11] K. Ask, M. Rhodin, L.-M. Tamminen, E. Hernlund, and P. Haubro Andersen, "Identification of Body Behaviors and Facial Expressions Associated with Induced Orthopedic Pain in Four Equine Pain Scales," *Animals*, vol. 10, no. 11, p. 2155, Nov. 2020, doi: https://doi.org/10.3390/ani10112155.

[12] A. Hussein, Ibraheem H.. M, Sarah Ali Abdulkareem, Ryam Ali Zubaid, and Noor Thamer, "Multi-Level Fusion for Facial Expression Recognition in Human Behavior Identification," vol. 10, no. 2, pp. 108–121, Jan. 2023, doi: https://doi.org/10.54216/fpa.100210.

[13] J. Zhou, L. Liu, W. Wei, and J. Fan, "Network Representation Learning: From Preprocessing, Feature Extraction to Node Embedding," *arXiv (Cornell University)*, Jan. 2021, doi: https://doi.org/10.48550/arxiv.2110.07582.

[14] Shilpa Gite *et al.*, "Textual Feature Extraction Using Ant Colony Optimization for Hate Speech Classification," *Big data and cognitive computing*, vol. 7, no. 1, pp. 45–45, Mar. 2023, doi: https://doi.org/10.3390/bdcc7010045.

[15] M. N. Kartheek, M. V. N. K. Prasad, and R. Bhukya, "Texture based feature extraction using symbol patterns for facial expression recognition," *Cognitive Neurodynamics*, Jun. 2022, doi: https://doi.org/10.1007/s11571-022-09824-z.

[16] G. Lei, X. Li, J. Zhou, and X. Gong, "Geometric feature based facial expression recognition using multiclass support vector machines," *IEEE*, Aug. 2009, doi: https://doi.org/10.1109/grc.2009.5255106.

[17] Reza abdi payamani, "Object Tracking based on Compressive Sensing Using Gabor Filters," *Power System Technology*, vol. 48, no. 1, pp. 2503–2513, Jun. 2024, doi: https://doi.org/10.52783/pst.521.

[18] Farzam Kharaji Nezhadian and S. Rashidi, "Palmprint verification based on textural features by using Gabor filters based GLCM and wavelet," *IEEE*, Mar. 2017, doi: https://doi.org/10.1109/csiec.2017.7940164.

[19] M. Sajjad *et al.*, "A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines," *Alexandria Engineering Journal*, vol. 68, pp. 817–840, Apr. 2023, doi: https://doi.org/10.1016/j.aej.2023.01.017.

[20] Y. ELsayed, A. ELSayed, and M. A. Abdou, "An automatic improved facial expression recognition for masked faces," *National Library of Medicine*, vol. 35, no. 20, pp. 14963–14972, Apr. 2023, doi: https://doi.org/10.1007/s00521-023-08498-w.

[21] Y.-C. Lin, T.-J. Cai, M.-H. Yen, C.-C. Chen, G.-S. Huang, and C.-Y. Peng, “Automatic Feeding System with High Accuracy Intelligent Product Defection Function,” *IEEE*, Jul. 2023, doi: https://doi.org/10.1109/coins57856.2023.10189198.

[22] Pratham Yashwante, Y. Patil, K. Nadar, and Anindita Khade, “Comparative Analysis of Meta-heuristic Feature Selection and Feature Extraction Approaches for Enhanced Chronic Kidney Disease Prediction,” *IEEE*, Mar. 2024, doi: https://doi.org/10.1109/iatmsi60426.2024.10502980.

[23] Alessio Santiccioli, M. Mercandelli, A.L. Lacaita, C. Samori, and S. Levantino, “A 1.6-to-3.0-GHz Fractional-${N}$ MDLL With a Digital-to-Time Converter Range-Reduction Technique Achieving 397-fs Jitter at 2.5-mW Power,” *IEEE Journal of Solid-state Circuits*, vol. 54, no. 11, pp. 3149–3160, Nov. 2019, doi: https://doi.org/10.1109/jssc.2019.2941259.

[24] K. Kousalya *et al.*, “Group Emotion Detection using Convolutional Neural Network,” *IEEE*, Feb. 2023, doi: https://doi.org/10.1109/otcon56053.2023.10113900.

[25] E. Lyakso, Ruban Nersisson, Olga Frolova, and M. A. Mekala, “The children’s emotional speech recognition by adults: Cross-cultural study on Russian and Tamil language,” *PLOS ONE*, vol. 18, no. 2, pp. e0272837–e0272837, Feb. 2023, doi: https://doi.org/10.1371/journal.pone.0272837.

[26] M. Ventura *et al.*, “Investigating the impact of disposable surgical face-masks on face identity and emotion recognition in adults with autism spectrum disorder,” *Autism Research*, Mar. 2023, doi: https://doi.org/10.1002/aur.2922.

[27] F. Mohammadi, F. Cheraghi, S. Khazaei, M. Seyedi, M. Rezaei, and F. Mirzaie, “Educational Facial Emotion Recognition in Children With Autism Spectrum Disorder: A Clinical Trial Study,” *Iranian Rehabilitation Journal*, vol. 20, no. 4, pp. 579–588, Dec. 2022, doi: https://doi.org/10.32598/irj.20.4.1734.1.

[28] Samaneh Madanian *et al.*, “Automatic Speech Emotion Recognition Using Machine Learning: Digital Transformation of Mental Health,” *AIS Electronic Library (AISeL)*, 2022. https://aisel.aisnet.org/pacis2022/45 (accessed Sep. 22, 2024).

[29] E. Nagy, L. Prentice, and T. Wakeling, “Atypical Facial Emotion Recognition in Children with Autism Spectrum Disorders: Exploratory Analysis on the Role of Task Demands,” *Perception*, vol. 50, no. 9, p. 030100662110381, Aug. 2021, doi: https://doi.org/10.1177/03010066211038154.

[30] C. J. Smith *et al.*, “Implementing the Get SET Early Model in a Community Setting to Lower the Age of ASD Diagnosis,” *Journal of Developmental & Behavioral Pediatrics*, vol. 43, no. 9, pp. 494–502, Dec. 2022, doi: https://doi.org/10.1097/dbp.0000000000001130.

[31] P. Washington *et al.*, “Crowdsourced privacy-preserved feature tagging of short home videos for machine learning ASD detection,” *Scientific Reports*, vol. 11, no. 1, Apr. 2021, doi: https://doi.org/10.1038/s41598-021-87059-4.

[32] A. Okewole, V.-R. Bourque, S. Jacquemont, V. Warrier, and S. Baron-Cohen, “Modelling Co-Occurring Mental Health Conditions Among Autistic Individuals Using Polygenic Scores,” BJPsych Open, vol. 10, no. S1, pp. S69–S69, Jun. 2024, doi: https://doi.org/10.1192/bjo.2024.222.

[33] M. Casseus, W. J. Kim, and D. B. Horton, “Prevalence and treatment of mental, behavioral, and developmental disorders in children with co-occurring autism spectrum disorder and attention-deficit/hyperactivity disorder: A population-based study,” *Autism Research*, Jan. 2023, doi: https://doi.org/10.1002/aur.2894.

[34] X. Cao and J. Cao, “Commentary: Machine learning for autism spectrum disorder diagnosis – challenges and opportunities – a commentary on Schulte-Rüther et al. (2022),” *Journal of Child Psychology and Psychiatry*, Feb. 2023, doi: https://doi.org/10.1111/jcpp.13764.

[35] K. Armstrong and S. W. Duvall, “Introductory editorial to the special issue: Assessment and diagnosis of autism spectrum disorder (ASD) and related clinical decision making in neuropsychological practice,” *The Clinical Neuropsychologist*, pp. 1–5, Jun. 2022, doi: https://doi.org/10.1080/13854046.2022.2085629.

[36] P. Esther Rani and S. Velmurugan, “Behavioral Analysis of Students by Integrated Radial Curvature and Facial Action Coding System using DCNN,” *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Mar. 2022, doi: https://doi.org/10.1109/icaccs54159.2022.9785056.

[37] M. Karnati, A. Seal, D. Bhattacharjee, A. Yazidi, and O. Krejcar, “Understanding Deep Learning Techniques for Recognition of Human Emotions Using Facial Expressions: A Comprehensive Survey,” *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–31, 2023, doi: https://doi.org/10.1109/TIM.2023.3243661.

[38] A. Kumar, None Navdeep, A. Kumar, Smruti Rekha Swain, I. Gupta, and Ashutosh Kumar Singh, “A Haar-Cascading and Deep Learning Driven Enhanced Mechanism on Cloud Platform for Real-Time Facial Emotion Detection,” *IEEE*, Jun. 2023, doi: https://doi.org/10.1109/icepe57949.2023.10201572.

[39] V. Upadhyay and D. Kotak, “A Review on Different Facial Feature Extraction Methods for Face Emotions Recognition System,” *2020 Fourth International Conference on Inventive Systems and Control (ICISC)*, Jan. 2020, doi: https://doi.org/10.1109/icisc47916.2020.9171172.

[40] F. M. Talaat, Z. H. Ali, R. R. Mostafa, and N. El-Rashidy, "Real-time facial emotion recognition model based on kernel autoencoder and convolutional neural network for autism children," *Soft Computing*, Jan. 2024, doi: https://doi.org/10.1007/s00500-023-09477-y.

[41] J. R. I. Coleman, K. J. Lester, R. Keers, M. R. Munafò, G. Breen, and T. C. Eley, "Genome-wide association study of facial emotion recognition in children and association with polygenic risk for mental health disorders," American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, vol. 174, no. 7, pp. 701–711, Jun. 2017, doi: https://doi.org/10.1002/ajmg.b.32558.

[42] G. Castellano, B. De Carolis, and N. Macchiarulo, "Automatic facial emotion recognition at the COVID-19 pandemic time," *Multimedia Tools and Applications*, Oct. 2022, doi: https://doi.org/10.1007/s11042-022-14050-0.

[43] S. Gokulakrishnan, P. Chakrabarti, Bui Thanh Hung, and S. Siva Shankar, "An optimized facial recognition model for identifying criminal activities using deep learning strategy," *International Journal of Information Technology*, vol. 15, no. 7, pp. 3907–3921, Sep. 2023, doi: https://doi.org/10.1007/s41870-023-01420-6.

[44] F. M. Talaat, "Real-time facial emotion recognition system among children with autism based on deep learning and IoT," *Neural Computing and Applications*, Mar. 2023, doi: https://doi.org/10.1007/s00521-023-08372-9.

[45] Benisha S. and Mirnalinee T.T., "Human Facial Emotion Recognition using Deep Neural Networks," *The international Arab journal of information technology*, vol. 20, no. 3, Jan. 2023, doi: https://doi.org/10.34028/iajit/20/3/2.

[46] N. Mehendale, "Facial emotion recognition using convolutional neural networks (FERC)," *SN Applied Sciences*, vol. 2, no. 3, Feb. 2020, doi: https://doi.org/10.1007/s42452-020-2234-1.

[47] M. Mukhiddinov, O. Djuraev, F. Akhmedov, A. Mukhamadiyev, and J. Cho, "Masked Face Emotion Recognition Based on Facial Landmarks and Deep Learning Approaches for Visually Impaired People," *Sensors*, vol. 23, no. 3, p. 1080, Jan. 2023, doi: https://doi.org/10.3390/s23031080.

[48] H.-C. Chu, W. W.-J. Tsai, M.-J. Liao, and Y.-M. Chen, "Facial emotion recognition with transition detection for students with high-functioning autism in adaptive e-learning," *Soft Computing*, vol. 22, no. 9, pp. 2973–2999, Apr. 2017, doi: https://doi.org/10.1007/s00500-017-2549-z.

[49] I. A. Ahmed *et al.*, "Eye Tracking-Based Diagnosis and Early Detection of Autism Spectrum Disorder Using Machine Learning and Deep Learning Techniques," *Electronics*, vol. 11, no. 4, p. 530, Feb. 2022, doi: https://doi.org/10.3390/electronics11040530.

[50] O. K. Oyedotun, G. G. Demisse, A. El, Djamila Aouada, and Bjorn Ottersten, "Facial Expression Recognition via Joint Deep Learning of RGB-Depth Map Latent Representations," *IEEE*, Oct. 2017, doi: https://doi.org/10.1109/iccvw.2017.374.

[51] S. Xu, X. Zhang, Y. Wu, F. Wei, and M. Zhou, "Unsupervised Extractive Summarization by Pre-training Hierarchical Transformers," *arXiv (Cornell University)*, Jan. 2020, doi: https://doi.org/10.48550/arxiv.2010.08242.

[52] R. Lubbers, M. F. van Essen, C. van Kooten, and L. A. Trouw, "Production of complement components by cells of the immune system," *Clinical & Experimental Immunology*, vol. 188, no. 2, pp. 183–194, Mar. 2017, doi: https://doi.org/10.1111/cei.12952.

[53] M. Liu, D. Yao, Z. Liu, J. Guo, and J. Chen, "An Improved Adam Optimization Algorithm Combining Adaptive Coefficients and Composite Gradients Based on Randomized Block Coordinate Descent," *Computational Intelligence and Neuroscience*, vol. 2023, pp. 1–14, Jan. 2023, doi: https://doi.org/10.1155/2023/4765891.

[54] H. Kohli, J. Agarwal, and M. Kumar, "An improved method for text detection using Adam optimization algorithm," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 230–234, Jun. 2022, doi: https://doi.org/10.1016/j.gltp.2022.03.028.

[55] N. Yoshida, S. Nakakita, and M. Imaizumi, "Effect of Random Learning Rate: Theoretical Analysis of SGD Dynamics in Non-Convex Optimization via Stationary Distribution," *arXiv (Cornell University)*, Jun. 2024, doi: https://doi.org/10.48550/arxiv.2406.16032.

[56] Vankadhara Rajyalakshmi and Kuruva Lakshmanna, "Detection of car parking space by using Hybrid Deep DenseNet Optimization algorithm," *International Journal of Network Management*, vol. 34, no. 1, Apr. 2023, doi: https://doi.org/10.1002/nem.2228.

[57] X. Yue and Q. Liu, "Improved FunkSVD Algorithm Based on RMSProp," *Journal of Circuits, Systems and Computers*, vol. 31, no. 08, Jan. 2022, doi: https://doi.org/10.1142/s0218126622501390.

[58] M. Huk, "Stochastic Optimization of Contextual Neural Networks with RMSprop," *Lecture notes in computer science*, pp. 343–352, Jan. 2020, doi: https://doi.org/10.1007/978-3-030-42058-1_29.

[59] K. D. Erickson and A. B. Smith, "Accounting for imperfect detection in data from museums and herbaria when modeling species distributions: combining and contrasting data-level versus model-level bias correction," *Ecography*, vol. 44, no. 9, pp. 1341–1352, Jul. 2021, doi: https://doi.org/10.1111/ecog.05679.

[60] T. Berg and P. N. Belhumeur, “POOF: Part-Based One-vs.-One Features for Fine-Grained Categorization, Face Verification, and Attribute Estimation,” *IEEE Xplore*, Jun. 01, 2013. https://ieeexplore.ieee.org/document/6618972 (accessed Apr. 11, 2023).

[61] R. Annamalai, Sudharson S, and Kolla Gnapika Sindhu, “Facial Matching and Reconstruction Techniques in Identification of Missing Person Using Deep Learning,” *IEEE*, Aug. 2023, doi: https://doi.org/10.1109/indiscon58499.2023.10270804.

[62] P. A. Chitale, K. Y. Kekre, H. R. Shenai, R. Karani, and J. P. Gala, “Pothole Detection and Dimension Estimation System using Deep Learning (YOLO) and Image Processing,” *IEEE Xplore*, Nov. 01, 2020. https://ieeexplore.ieee.org/abstract/document/9290547 (accessed Dec. 15, 2022).

[63] E. S. Salama, R. A. El-Khoribi, M. E. Shoman, and M. A. Wahby Shalaby, “A 3D-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition,” *Egyptian Informatics Journal*, Aug. 2020, doi: https://doi.org/10.1016/j.eij.2020.07.005.

[64] H.-D. Nguyen, S. Yeom, G.-S. Lee, H.-J. Yang, I.-S. Na, and S.-H. Kim, “Facial Emotion Recognition Using an Ensemble of Multi-Level Convolutional Neural Networks,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 33, no. 11, p. 1940015, Oct. 2019, doi: https://doi.org/10.1142/s0218001419400159.

[65] X. Wang, M. Peng, L. Pan, M. Hu, C. Jin, and F. Ren, “Two-level Attention with Two-stage Multi-task Learning for Facial Emotion Recognition,” *arXiv (Cornell University)*, Jan. 2018, doi: https://doi.org/10.48550/arxiv.1811.12139.

[66] L. Chen, M. Zhou, W. Su, M. Wu, J. She, and K. Hirota, “Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction,” *Information Sciences*, vol. 428, pp. 49–61, Feb. 2018, doi: https://doi.org/10.1016/j.ins.2017.10.044.

[67] A. K. Hassan and S. N. Mohammed, “A novel facial emotion recognition scheme based on graph mining,” *Defence Technology*, Dec. 2019, doi: https://doi.org/10.1016/j.dt.2019.12.006.

[68] A. Gupta, S. Arunachalam, and R. Balakrishnan, “Deep self-attention network for facial emotion recognition,” *Procedia Computer Science*, vol. 171, pp. 1527–1534, 2020, doi: https://doi.org/10.1016/j.procs.2020.04.163.

[69] D. Jiang *et al.*, “A probability and integrated learning based classification algorithm for high-level human emotion recognition problems,” *Measurement*, vol. 150, p. 107049, Jan. 2020, doi: https://doi.org/10.1016/j.measurement.2019.107049.

[70] 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, Apr. 2020, doi: https://doi.org/10.1016/j.bspc.2020.101867.

[71] M. Rescigno, M. Spezialetti, and S. Rossi, "Personalized models for facial emotion recognition through transfer learning," *Multimedia Tools and Applications*, vol. 79, no. 47–48, pp. 35811–35828, Aug. 2020, doi: https://doi.org/10.1007/s11042-020-09405-4.

[72] X. Liu, X. Cheng, and K. Lee, "GA-SVM based Facial Emotion Recognition using Facial Geometric Features," *IEEE Sensors Journal*, pp. 1–1, 2020, doi: https://doi.org/10.1109/jsen.2020.3028075.

[73] G. Yolcu *et al.*, "Deep learning-based facial expression recognition for monitoring neurological disorders," *IEEE Xplore*, Nov. 01, 2017. https://ieeexplore.ieee.org/abstract/document/8217907/

[74] Z. Gao, Y. Li, Y. Yang, X. Wang, N. Dong, and H.-D. Chiang, "A GPSO-optimized convolutional neural networks for EEG-based emotion recognition," *Neurocomputing*, vol. 380, pp. 225–235, Mar. 2020, doi: https://doi.org/10.1016/j.neucom.2019.10.096.

[75] D. Wu, J. Zhang, and Q. Zhao, "Multimodal Fused Emotion Recognition About Expression-EEG Interaction and Collaboration Using Deep Learning," *IEEE Access*, vol. 8, pp. 133180–133189, Jan. 2020, doi: https://doi.org/10.1109/access.2020.3010311.

[76] Q. T. Ngoc, S. Lee, and B. C. Song, "Facial Landmark-Based Emotion Recognition via Directed Graph Neural Network," *Electronics*, vol. 9, no. 5, p. 764, May 2020, doi: https://doi.org/10.3390/electronics9050764.

[77] I. Bendjoudi, F. Vanderhaegen, D. Hamad, and F. Dornaika, "Multi-label, multi-task CNN approach for context-based emotion recognition," *Information Fusion*, Nov. 2020, doi: https://doi.org/10.1016/j.inffus.2020.11.007.

[78] C. Tsangouri, W. Li, Z. Zhu, F. Abtahi, and T. Ro, "An interactive facial-expression training platform for individuals with autism spectrum disorder," *2016 IEEE MIT Undergraduate Research Technology Conference (URTC)*, Nov. 2016, doi: https://doi.org/10.1109/urtc.2016.8284067.

[79] J. Manfredonia *et al.*, "Automatic Recognition of Posed Facial Expression of Emotion in Individuals with Autism Spectrum Disorder," *Journal of Autism and Developmental Disorders*, vol. 49, no. 1, pp. 279–293, Oct. 2018, doi: https://doi.org/10.1007/s10803-018-3757-9.

[80] D. N. McIntosh, A. Reichmann-Decker, P. Winkielman, and J. L. Wilbarger, "When the social mirror breaks: deficits in automatic, but not voluntary, mimicry of emotional facial expressions in autism," *Developmental Science*, vol. 9, no. 3, pp. 295–302, May 2006, doi: https://doi.org/10.1111/j.1467-7687.2006.00492.x.

[81] I. Gordon, M. D. Pierce, M. S. Bartlett, and J. W. Tanaka, “Training Facial Expression Production in Children on the Autism Spectrum,” *Journal of Autism and Developmental Disorders*, vol. 44, no. 10, pp. 2486–2498, Apr. 2014, doi: https://doi.org/10.1007/s10803-014-2118-6.

[82] Tang *et al.*, “CFEW: A Large-Scale Database for Understanding Child Facial Expression in Real World,” *IEEE Transactions on Affective Computing*, vol. 15, no. 3, pp. 990–1003, Jul. 2024, doi: https://doi.org/10.1109/taffc.2023.3313782.

[83] Tang, Chuangao, WenmingZheng, Yuan Zong, Zhen Cui, Nana Qiu, Simeng Yan, and XiaoyanKe. "Automatic smile detection of infants in mother-infant interaction via CNN-based feature learning." In Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data, pp. 35-40. 2018.

[84] Alamgir, F.M. and Alam, M.S. (2022). A Novel Deep Learning-based Bidirectional Elman Neural Network for Facial Emotion Recognition. International Journal of Pattern Recognition and Artificial Intelligence. doi:https://doi.org/10.1142/s0218001422520164.

[85] Vignesh, S., Savithadevi, M., Sridevi, M. and Sridhar, R. (2023). A novel facial emotion recognition model using segmentation VGG-19 architecture. International Journal of Information Technology. doi:https://doi.org/10.1007/s41870-023-01184-z.

[86] Rahkar Farshi, T. (2020). Battle royale optimization algorithm. Neural Computing and Applications. doi:https://doi.org/10.1007/s00521-020-05004-4.

[87] Pauline, O., Ong Kok Meng and Sia Chee Kiong (2017). An improved flower pollination algorithm with chaos theory for function optimization. AIP conference proceedings. doi:https://doi.org/10.1063/1.4995922.

[88] Jia, W., Zhao, D., Shen, T., Tang, Y. and Zhao, Y. (2014). Study on Optimized Elman Neural Network Classification Algorithm Based on PLS and CA. Computational Intelligence and Neuroscience, 2014, pp.1–13. doi:https://doi.org/10.1155/2014/724317.

[89] Ai, H., Wu, S., Gao, H., Zhao, L., Yang, C. and Zhang, Y. (2012). Temperature distribution analysis of tissue water vaporization during microwave ablation: Experiments and simulations. International Journal of Hyperthermia, 28(7), pp.674–685. doi:https://doi.org/10.3109/02656736.2012.710769.

[90] Singh, A., Chodankar, S. and Suvarna, A. (2340). AUDIO FEATURE EXTRACTION TOOLS. [online] International Research Journal of Modernization in Engineering Technology and Science @International Research Journal of Modernization in Engineering, pp.2582–5208. Available at: https://www.irjmets.com/uploadedfiles/paper/volume3/issue_4_april_2021/9248/1628083376.pdf [Accessed 28 Aug. 2024].

[91] Feng, T. and Yang, S. (2018). Speech Emotion Recognition Based on LSTM and Mel Scale Wavelet Packet Decomposition. Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence. doi:https://doi.org/10.1145/3302425.3302444.

[92] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Communications of the ACM, 60(6), pp.84–90.

[93] Nie, X., Zhou, X., Tong, T., Lin, X., Wang, L., Zheng, H., Li, J., Xue, E., Chen, S., Zheng, M., Chen, C. and Du, M. (2022). N-Net: A novel dense fully convolutional neural network for thyroid nodule segmentation. Frontiers in Neuroscience, 16. doi:https://doi.org/10.3389/fnins.2022.872601.

[94] Abdulhussien, W.R., El Abbadi, N.K. and Gaber, A.M. (2021). Hybrid Deep Neural Network for Facial Expressions Recognition. Indonesian Journal of Electrical Engineering and Informatics (IJEEI), 9(4). doi:https://doi.org/10.52549/ijeei.v9i4.3425.

[95] Bisogni, C., Castiglione, A., Hossain, S., Narducci, F. and Umer, S. (2022). Impact of Deep Learning Approaches on Facial Expression Recognition in Healthcare Industries. IEEE Transactions on Industrial Informatics, [online] 18(8), pp.5619–5627. doi:https://doi.org/10.1109/TII.2022.3141400.