

CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY, ISLAMABAD



Multi-model CNN based Facial Emotion Recognition for Individuals with Autism Spectrum Disorder

by

Hira Iqbal

A thesis submitted in partial fulfillment for the
degree of Master of Science

in the

Faculty of Computing

Department of Computer Science

2026

Copyright © 2026 by Hira Iqbal

All rights reserved. No part of this thesis may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, by any information storage and retrieval system without the prior written permission of the author.



CERTIFICATE OF APPROVAL

Multi-model CNN based Facial Emotion Recognition for Individuals with Autism Spectrum Disorder.

by

Hira Iqbal

(MAI241003)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Muhammad Nazir	IIUI, Islamabad
(b)	Internal Examiner	Dr. M. Masroor Ahmed	CUST, Islamabad

Dr. Muhammad Siraj Rathore

Thesis Supervisor

September, 2026

Dr. Nadeem Anjum

Head

Dept. of Artificial Intelligence

September, 2026

Dr. M. Abdul Qadir

Dean

Faculty of Computing

September, 2026

Author's Declaration

I, **Hira Iqbal** hereby state that my MS thesis titled “**Multi-model CNN-based Facial Emotion Recognition for Individuals with Autism Spectrum Disorder**” is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time, if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.



(Hira Iqbal)

Reg No:MAI241003

Plagiarism Undertaking

I solemnly declare that the research work presented in this thesis titled “**Multi-model CNN-based Facial Emotion Recognition for Individuals with Autism Spectrum Disorder**”, is solely my research work with no significant contribution from any other person. Any small contribution or help received has been duly acknowledged, and I have written the complete thesis myself.

I understand the zero-tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I, as an author of the above-titled thesis, declare that no portion of my thesis has been plagiarised and any material used as a reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after the award of the MS Degree, the University reserves the right to with- draw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.



(Hira Iqbal)

Reg No:MAI241003

Acknowledgement

All worship and praise is for ALLAH (S.W.T), the creator of the whole world. First and foremost, I would like to thank Him for providing me the strength, knowledge and blessings to complete this research work. Secondly, thanks to my respected supervisor, Dr Muhammad Siraj Rathore, for his assistance, valuable time and guidance. I sincerely thank him for his support, encouragement and advice in the research area. He enabled me to develop an understanding of the subject. He has taught me, consciously and unconsciously, how good experimental work is carried out. Sir, you will always be part of my prayers throughout my life. I would also like to thank all research group members for their comments and feedback on my research work.

I am highly beholden to my family for their assistance, support (moral as well as financial), and encouragement throughout the completion of this Master of Science degree. This is all due to the love they show me in every moment of my life. No words can ever be sufficient for my gratitude to my parents. I hope I have met my parents' high expectations.

I dedicate my dissertation work to my Parents, Family, Teachers and Friends. A special feeling of gratitude is for my loving mother, sister and brothers for their love, endless support and Duas. I owe everything to them. A special thanks to my supervisor for believing in me and inspiring me to work harder and achieve greater things in life. This work is also a dedication to all the women who strive hard for their careers regardless of the countless society pressures and discouragements.

I pray to ALLAH (S.W.T) that He bestow me with true success in all fields in both worlds and shower His blessed knowledge upon me to better all Muslims and whole Mankind. Ameen!

(Hira Iqbal)

Abstract

Facial Emotion Recognition (FER) can play an important role in supporting communication, social interaction, and emotional development among individuals with Autism Spectrum Disorder (ASD). However, existing FER systems often face challenges related to limited and imbalanced datasets, as well as differences in facial expressions between individuals with ASD and neurotypical individuals. These challenges make it difficult to achieve consistently high recognition accuracy and ensure the generalizability and robustness of FER models across different datasets, populations, and experimental settings. To address these challenges, this study develops a deep learning-based FER framework specifically targeting ASD-related facial expressions. A cleaned hybrid dataset was constructed by integrating the FER-2013 dataset with a facial emotion dataset of autistic children developed by Fatima Tallat. The integration was intended to improve class representation, increase data diversity, and provide a more comprehensive basis for evaluating FER models in ASD-related settings. The study investigates the effectiveness of transfer learning-based Convolutional Neural Networks (CNNs) for facial emotion classification and evaluates the impact of the proposed hybrid dataset on model performance. Five state-of-the-art pretrained architectures, namely VGG16, ResNet50, DenseNet121, MobileNetV2, and Xception, were implemented and trained separately on the FER-2013 dataset, the Fatma Talat ASD-specific dataset, and the proposed hybrid dataset. Using the same CNN architectures and consistent experimental settings across all three datasets enables a fair comparison of model performance and provides insights into whether integrating the datasets improves recognition accuracy, robustness, and generalizability across different data settings. Among the evaluated models, VGG16 trained on the proposed hybrid dataset achieved the best performance, attaining an accuracy of 91%. Based on the performance of the individual models, the three best-performing CNN architectures were subsequently selected and combined to develop a multimodel CNN configuration. The model fusion approach was designed to exploit the complementary learning capabilities of the selected architectures and further improve emotion classification accuracy and prediction consistency. Among the individual CNN

models trained on the proposed hybrid dataset, ResNet50, Xception, and VGG16 demonstrated the best overall performance and were therefore selected for further multimodel experimentation. Different combinations of these top-performing architectures were subsequently evaluated to investigate the effectiveness of model fusion. The results showed that the ResNet50–VGG16 combination achieved the best performance, attaining an accuracy of 99% on the hybrid dataset, compared with 91% achieved by the best-performing individual model, VGG16. Overall, the findings demonstrate that combining a carefully constructed hybrid dataset with transfer learning and multimodel CNN fusion can improve the accuracy, robustness, and generalizability of FER systems for individuals with ASD.

Contents

Author’s Declaration	iii
Plagiarism Undertaking	iv
Acknowledgement	v
Abstract	vi
List of Figures	xii
List of Tables	xiv
Abbreviations	xv
1 Introduction	1
1.1 Autism Spectrum Disorder and Emotional Communication	2
1.2 Facial Emotion Recognition and Its Role in ASD	2
1.3 Machine Learning and Deep Learning in FER	3
1.4 Integration of Deep Learning and Facial Emotion Recognition	3
1.4.1 Transfer Learning in Facial Emotion Recognition	4
1.4.2 Common Deep Learning Architectures for FER	5
1.4.2.1 VGG16	5
1.4.2.2 ResNet50	5
1.4.2.3 DenseNet 121	5
1.4.2.4 Xception	6
1.4.2.5 MobileNet	6
1.4.3 Commonly Used Datasets in FER	7
1.4.3.1 FER-2013 Dataset	7
1.4.3.2 RAF-DB Dataset	8
1.4.3.3 CK+ Dataset	8
1.4.3.4 Fatima Talaat Autism Dataset	8
1.5 Problem Statement	8
1.6 Research Objectives	9

1.7	Research Questions	9
1.8	Rationale and Significance of Selected Research Topic	10
1.9	Research Methodology	10
1.9.1	Problem Formulation	11
1.9.2	Dataset Selection	11
1.9.3	Data Preprocessing	11
1.9.4	Model Selection and Architecture Design	11
1.9.5	Transfer Learning Implementation	12
1.9.6	Model Training and Validation	12
1.9.7	Performance Evaluation	12
1.9.8	Cross-Dataset and Cross-Model Comparison	12
1.9.9	Hybrid Model Development	13
1.9.10	Outcome Analysis and Insights	13
1.10	Important Findings	13
1.11	Summary and Thesis Outline	14
2	Literature Review	15
2.1	Introduction	15
2.2	Studies Introducing a Facial Emotion Recognition System in Neurotypical Individual	16
2.3	Studies Based on Emotion Recognition in Autistic Children	18
2.3.1	Using Machine learning approach	18
2.3.2	Using a Deep learning approach	19
2.4	Comparison of the Articles	24
2.5	Identified Research Gap	24
2.6	Summary	26
3	Methodology	27
3.1	Introduction	27
3.2	Objective	28
3.3	Model Selection	28
3.4	Model Architecture	29
3.5	Model Training	30
3.5.1	Employing Pretrained Architectures	30
3.5.2	Application of Transfer Learning and Feature Extraction	30
3.5.3	Custom Classification Head	31
3.5.4	Model Compilation	31
3.5.5	Training Procedure	32
3.5.6	Model Saving	32
3.6	Model Training on Individual Dataset	32
3.7	FER13	33
3.8	Fatima Tallat's Dataset for Autistic Children	33
3.8.1	Dataset Preprocessing	33
3.9	Noise Removal and Filtering:	34
3.10	Model Training on Hybrid Dataset	35

3.10.1	Hybrid Dataset	36
3.10.2	Handling Class Imbalance for Hybrid Dataset	37
3.11	Multi-Model Training on Hybrid Dataset	38
3.12	Multi-model Model Architecture	39
3.12.1	Frozen Layers	41
3.13	Performance Metrics	43
3.13.1	Accuracy	43
3.13.2	Precision	43
3.13.3	Recall (also referred to as Sensitivity)	44
3.13.4	F1-Score	44
3.13.5	Confusion Matrix	44
3.13.6	A Training and Validation Curve	45
3.14	Tools and Environment	45
3.15	Summary	46
4	Results and Discussions	47
4.1	Implementation Environment	48
4.2	Model Implementation on Fatima tallat's autistic children dataset	49
4.2.1	Xception	49
4.2.1.1	Per-Class Metrics	50
4.2.2	ResNet	51
4.2.3	Vgg16	53
4.2.3.1	Per-Class Metrics	54
4.3	Performance metrics	55
4.3.1	Observation	55
4.4	Model Implementation on FER13 Dataset	56
4.4.1	VGG16 on FER13 Dataset	56
4.4.2	ResNet50 on FER13 Dataset	58
4.4.3	DenseNet121 on FER13 Dataset	59
4.4.4	MobileNetV2 on FER13 dataset	61
4.4.5	Xception on FER13 dataset	62
4.5	Comparative Results of FER13 Dataset	64
4.6	Facial Emotion Recognition , Hybrid Dataset	65
4.6.1	VGG16 on Hybrid Dataset	65
4.6.2	Xception on HYBRID Dataset	68
4.6.3	ResNet on HYBRID Dataset	70
4.6.4	Performance Comparizon	72
4.7	Multimodels on Hybrid Dataset	73
4.7.1	ResNet+VGG16 on HYBRID Dataset	73
4.7.2	ResNet+Xception on HYBRID Dataset	75
4.7.3	VGG16+Xception on HYBRID Dataset	78
4.7.4	Comparative Results on HYBRID Dataset	80
4.8	Chapter Summary	82

5	Conclusion and Future Work	83
5.1	Summary of Findings	84
5.1.1	Dataset Insights	84
5.1.2	Model Performance	84
5.1.3	Error Analysis	85
5.1.4	Findings for ASD-FER	86
5.2	Theoretical Findings	86
5.3	Practical Finding	87
5.4	Limitations of the Study	88
5.5	Future Research Directions	88
5.5.1	Model Explainability	88
5.5.2	Real-Time and Edge Deployment	89
5.5.3	Ethical and Social Dimensions	90
	Bibliography	91

List of Figures

1.1	Architecture for VGG16 layers, reproduced by [15]	5
1.2	ResNet-50 Layers Architecture, reproduced from [17]	6
1.3	Workflow of the densenet121 model, reproduced by [19].	6
1.4	Architecture for the xception model, reproduced by [20]	7
1.5	Architecture for MobileNet, reproduced by [22]	7
3.1	Model traing on FER13 and ASD Fatima Tallat, indivisual datasets	32
3.2	Model training mechanism for proposed system using hybrid dataset.	35
3.3	Bar Chat for Emotion Distribution Across Hybrid Dataset showing class inbalance.	36
3.4	Multi-model based training mechanism on hybrid dataset	38
3.5	Multi-Model CNN based Architecture	42
4.1	Accuracy and Loss curves of Xception on ASD FT dataset	49
4.2	Confusion matrix of the Austisitc Childern Emotions by Xception Model on ASD FT dataset	51
4.3	Accuracy and Loss curves of ResNet on ASD FT dataset	52
4.4	Confusion matrix of the Austisitc Childern Emotions by Xception Model by Xception Model on ASD FT dataset	53
4.5	Accuracy and Loss curves of VGG16 on ASD FT dataset	53
4.6	Confusion matrix of the Austisitc Childern Emotions by VGG16 Model by Xception Model on ASD FT dataset	54
4.7	VGG16 Training loss accuracy graph on FER13 dataset	56
4.8	Confusion matrix by VGG16 on FER13.	57
4.9	ResNet50 Training loss accuracy graph on FER13 dataset	58
4.10	Confusion matrix by ResNet on FER13.	59
4.11	DenseNet121 Training loss accuracy graph on FER13 dataset.	60
4.12	Confusion matrix by DenseNet on FER13	60
4.13	MobileNet ResNet50 Training loss accuracy graph on FER13 dataset	61
4.14	Confusion matrix by MobileNet on FER13.	62
4.15	Xception Training loss accuracy graph on FER13 dataset	63
4.16	Confusion matrix by Xception model on the FER13 dataset.	63
4.17	VGG16 Training loss accuracy graph on FER13 dataset	66
4.18	Confusion Matrix of vgg16 on hybrid dataset	67
4.19	Xception Model Accuracy and Loss on Hybrid Dataset	69
4.20	Xception Confusion Metric on Hybrid Dataset	69

4.21	ResNet Model Accuracy and Loss on Hybrid Dataset	70
4.22	ResNet Confusion Metric on Hybrid Dataset	71
4.23	ResNet + VGG16 Model Training Accuracy and Loss on Hybrid Dataset	73
4.24	ResNet + VGG16 Confusion Matrix on Hybrid Dataset	74
4.25	ResNet + Xception Model Training Accuracy and Loss on Hybrid Dataset	76
4.26	ResNet + Xception Classification Report on Hybrid Dataset	77
4.27	VGG16 + Xception Model Training Accuracy and Loss on Hybrid Dataset	78
4.28	VGG16 + Xception Classification Report on Hybrid Dataset	79

List of Tables

2.1	Summary of Related Work on Facial Emotion Recognition for Autism Spectrum Disorder	25
2.2	Most Relevant Studies from Table 2.1 Summary	26
3.1	Confusion matrix showing placement for samples	45
4.1	per-class Classification report for Xception on ASD FT dataset	51
4.2	Per-class Classification Report of Emotion Recognition of ResNet on ASD FT dataset	52
4.3	per-class Classification report for Vgg16 ASD FT dataset	55
4.4	Overall Performance Metric on ASD FT datasets	55
4.5	Overall Performance Metrics on FER13 dataset	64
4.6	Per-Class ,Classification report for Vgg16 on Hybrid dataset	68
4.7	Per-class classification report for xception on Hybrid dataset	70
4.8	Per-class Classification Report of Emotion Recognition of ResNet on Hybrid dataset	71
4.9	Performance Comparison of Different Models on Hybrid Dataset	72
4.10	ResNet + VGG16 Classification Report on Hybrid Dataset	75
4.11	ResNet + Xception Confusion Matrix on Hybrid Dataset	77
4.12	VGG16 + Xception Confusion Matrix on Hybrid Dataset	80
4.13	Performance Comparison of Different Models on Hybrid Dataset	80

Abbreviations

ASD	Autism Spectrum Disorder
AUC	Area Under the Curve
CNN	Convolutional Neural Network
DenseNet	Densely Connected Convolutional Networks
FACS	Facial Action Coding System
FER	Facial Emotion Recognition
FN	False Negatives
FP	False Positives
F1-Score	Harmonic Mean of Precision and Recall
GAP	Global Average Pooling
ImageNet	Large Visual Database used for Deep Learning Model Pre-training
MobileNet	Lightweight CNN Architecture for Mobile Devices
RBF	Radial Basis Function
ResNet	Residual Network
ROC	Receiver Operating Characteristic
TN	True Negatives
TP	True Positives
VGG16	Visual Geometry Group 16 (CNN Architecture)
Xception	Extreme Version of Inception CNN Architecture

Chapter 1

Introduction

Over the last several years, one has witnessed growing interest in developing smarter technologies capable of identifying facial expression in people with Autism Spectrum Disorder (ASD), with the view to enhancing the diagnostic and interventionale procedures [1]. The prevalence of Autism Spectrum Disorder (ASD) has increased steadily worldwide over the past two decades . Recent estimates indicate that approximately 1 in 100 children globally are affected by ASD, highlighting its growing public health and developmental significance. According to the latest data from the U.S. Centers for Disease Control and Prevention (CDC), approximately 1 in 31 children aged 8 years in the United States were identified with ASD in 2022, compared with 1 in 36 children in 2020 [2]. The Centres of Disease Control and Prevention (CDC) says that with the proper and early-identified and delivered depending on their needs, autistic individuals can radically enhance their social communication and adaptive performance [3]. But standard systems ASD facial emotion recognition (FER) systems, in general, cannot tend to identify the normal and discrete faces of emotions, which restricts their application in practice [4]. In response, state-of-the-art deep learning models, both with general and autism-specific datasets, are becoming increasingly common in use to enhance accuracy of FER and situational appropriateness in ASD situations. The following section will delve into the basis of facial expression recognition and explain the fundamental principles behind how humans and machines identify emotional cues

from facial features. It will discuss the psychological and computational foundations that support the interpretation of facial signals. Additionally, the section will explore how facial expression recognition has found meaningful application in autism studies, particularly in understanding social and emotional communication challenges.

1.1 Autism Spectrum Disorder and Emotional Communication

Individuals with ASD often experience serious difficulties with verbal and non-verbal interaction, as well as recognition and expression of facial emotions, which are essential concepts of social interaction [5]. These challenges have a significant impact on their involvement in regular social situations and might lead to isolation and low quality of life. Research has always shown that people with ASD have uncharacteristic facial emotion recognition and expressions where they do not always show emotions based on their intensity but also when and in what contexts [6]. Such distinctions cause significant difficulties in clinical assessments and interventions, including the accurate interpretation of emotional indicators to avert effective provision of behavioural assistance and contrivance of therapy.

1.2 Facial Emotion Recognition and Its Role in ASD

Facial Emotion Recognition (FER) technology has emerged in order to improve the understanding and social abilities of autistic individuals. Computer vision is used in FER systems to identify joy, sadness, anger, and surprise on the faces based on facial expressions in real-time systems [5]. However, most FER systems have been trained on datasets such as FER-2013 and CK+ which are based on samples of general populations and are not representative of people on the autism

spectrum [6]. As a result, such models are often unable to identify the delicate or less common manifestations common in ASD, decreasing their effectiveness in clinical and educational settings. This limitation can lead to inaccurate assessments and reduced reliability when these systems are used to support diagnosis or intervention. Consequently, it highlights the need for more specialized and inclusive models that are better trained to recognize the diverse expression patterns associated with ASD.

1.3 Machine Learning and Deep Learning in FER

Convolutional Neural Networks (CNNs) in machine learning have particularly advanced the accuracies of facial emotion recognition (FER). Modern models such as VGG16, ResNet, DenseNet, MobileNet, and Xception demonstrated excellent accuracy on conventional datasets [7].

The transfer learning method has enhanced the efficiency with which models are trained to allow finer-tuning of pre-trained models based on domain data with smaller samples. This is especially beneficial in the study of autism spectrum disorder (ASD) where the amount of available data is sometimes limited and not diverse. However, little has involved specific tailoring to ASD populations or evaluating them in terms of interpretability and real-life applicability in treatment settings.

1.4 Integration of Deep Learning and Facial Emotion Recognition

Facial Emotion Recognition (FER) has truly become one of the major components of affective computing, as it aids machines to interpret and respond to human emotions in accordance with facial expressions [8].

Traditional machine learning methods are less suitable for FER in autistic children because they often rely on manually extracted features such as facial landmarks, edges, and textures. These features may not effectively capture the subtle, atypical, and complex facial expressions of autistic children. ML models can also require extensive feature engineering and may have limited ability to learn high-level patterns directly from facial images[9].

Deep learning is more suitable because it can automatically learn and extract complex facial features directly from images without relying heavily on manual feature engineering. Models such as ResNet, VGG16, and Xception can identify subtle patterns and hierarchical features, making them more capable of recognising diverse and atypical expressions.

Deep learning also supports transfer learning, which is particularly useful when ASD-specific datasets are small. [6]. Transfer learning has also emerged as a useful method to fine-tune these pre-trained networks to FER tasks in particular when labelled data is limited. This saves on a lot of computational efforts and minimizes on training time yet giving excellent results [10][11][12]

1.4.1 Transfer Learning in Facial Emotion Recognition

In recent years, transfer learning has emerged as an important method in deep learning especially on tasks with limited annotated data, such as Facial Emotion Recognition (FER) using models trained on large datasets such as ImageNet [8].

Neural networks (CNNs) to specific tasks, and specifically fine-tune the upper layers, which makes this transfer learning especially beneficial in medical and psychological research where the amount of data to train on is significantly limited.

This approach reduces the need to train a deep neural network entirely from scratch and helps improve training efficiency and model stability. Furthermore, transfer learning can enable the model to learn task-specific patterns while retaining the general visual features acquired from large-scale datasets.[13]

1.4.2 Common Deep Learning Architectures for FER

Current facial expression recognition (FER) systems have utilised different deep convolutional neural network (CNN) architectures because of their exceptional ability to learn features: here some most well known model of deep learning architecture are discussed :

1.4.2.1 VGG16

VGG16 is a 16-layer deep CNN featuring small 3x3 filters and has achieved considerable fame because of its simplicity, as well as, the great performance across numerous image recognition tasks. It is usually used in FER due to its efficiency in retrieving layered facial expressions [14].

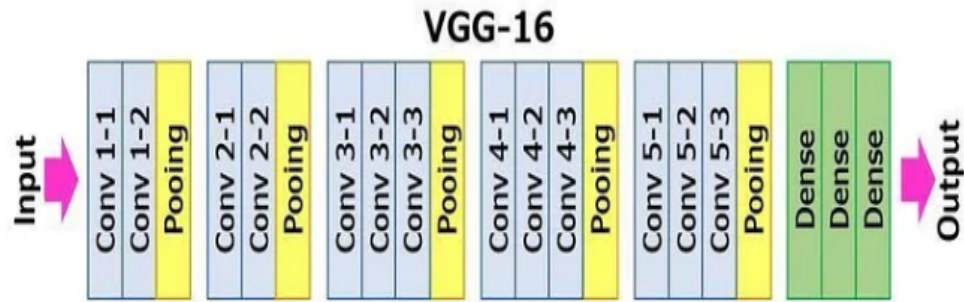


Figure 1.1: Architecture for VGG16 layers, reproduced by [15]

1.4.2.2 ResNet50

ResNet involves the use of shortcut connections, which are referred to as residual learning to effectively train deep networks. [16].

1.4.2.3 DenseNet 121

DenseNet builds connections between layers to all other layers, which promotes feature reuse and enhances the efficiency of gradient flow. This tier of connectivity is able to identify minor and intricate facial expressions [18].

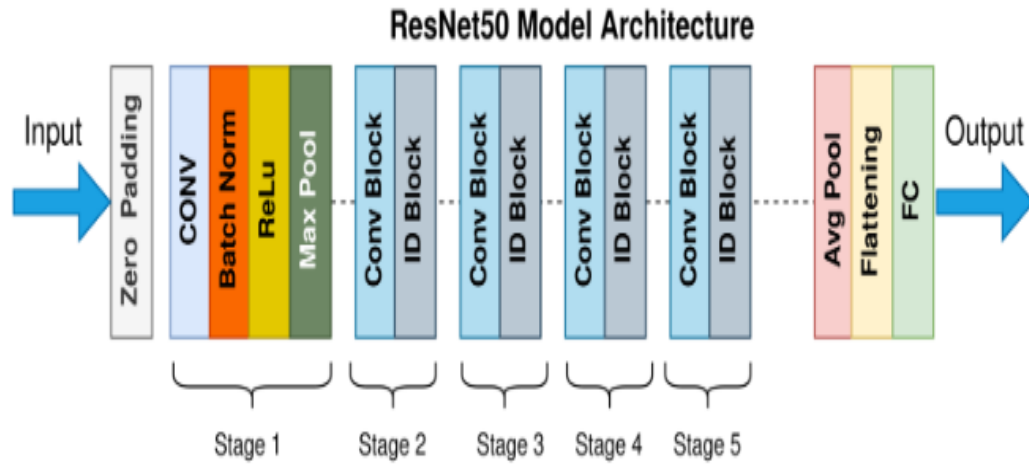


Figure 1.2: ResNet-50 Layers Architecture, reproduced from [17]

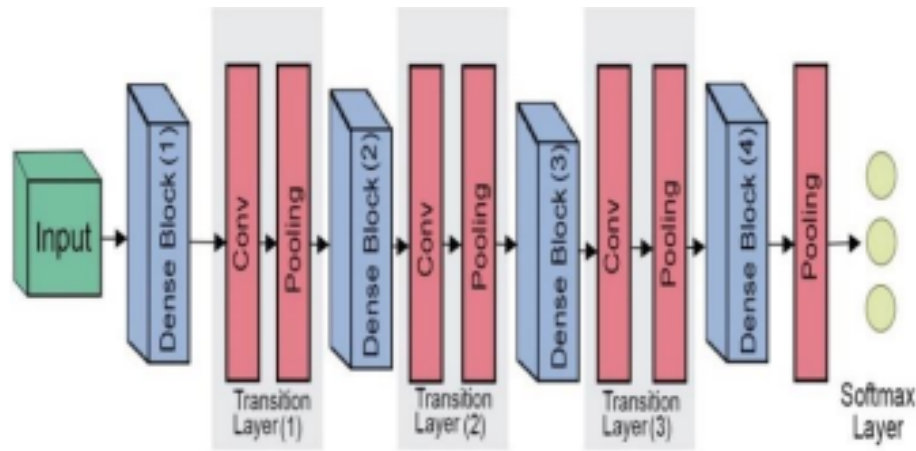


Figure 1.3: Workflow of the densenet121 model, reproduced by [19].

1.4.2.4 Xception

Xception applies depthwise separable convolutions to offer an efficient computation mechanism whilst preserving accuracy in representation during classification, hence is suitable in facial expression recognition on the fly or when embedded with other applications [20].

1.4.2.5 MobileNet

Depthwise separable convolution is also used in MobileNet to reduce computational complexity while maintaining effective feature extraction. Its lightweight

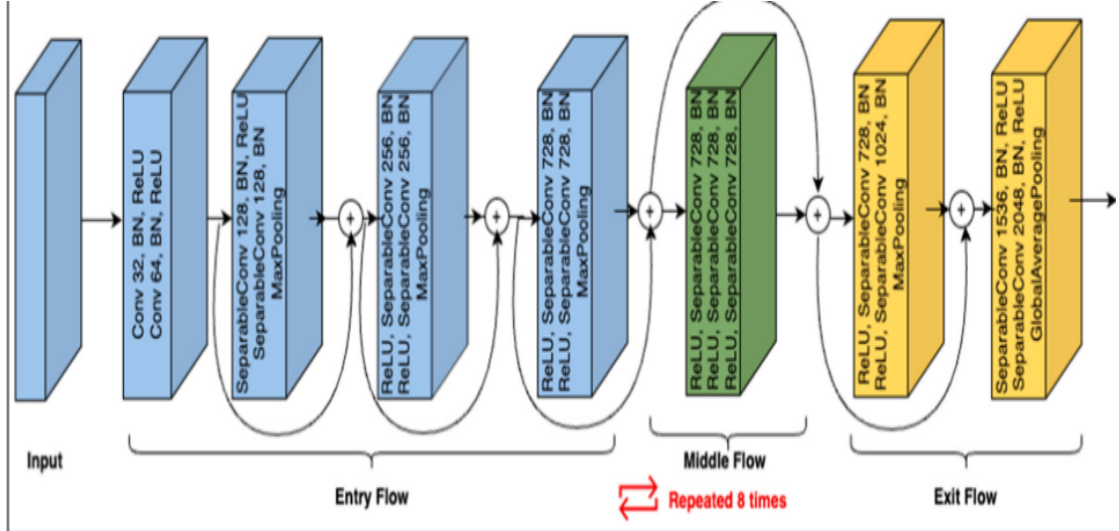


Figure 1.4: Architecture for the xception model, reproduced by [20]

architecture makes MobileNet suitable for low latency and resource constrained applications.[21].

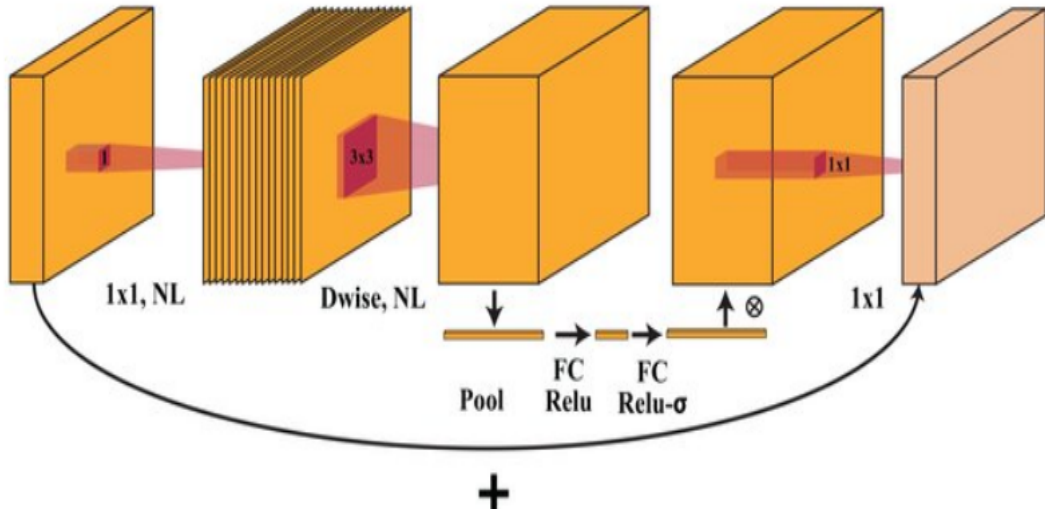


Figure 1.5: Architecture for MobileNet, reproduced by [22]

1.4.3 Commonly Used Datasets in FER

1.4.3.1 FER-2013 Dataset

FER-2013 is a popular facial expression recognition (FER) benchmark dataset which consists of 35,887 low-resolution grayscale (48x48 pixels) images, with 7 basic emotion categories [6]. Although it is common because it is relatively large

and is available, it has the problems that it is noisy when labelling and is not variably subtle in expressions. It is a generic dataset not ASD specific.

1.4.3.2 RAF-DB Dataset

RAF-DB contains over 29,000 face images which have been labelled using basic and compound affective labels. Its diverse poses, light and occlusion properties make it especially useful in evaluating FER models in realistic conditions [23]. It is a generic dataset not ASD specific.

1.4.3.3 CK+ Dataset

The Extended Cohn-Kanade (CK+) dataset is a generic facial emotion dataset, containing 593 sequences that capture facial expressions from neutral to peak emotional states. [24]. It is a generic dataset not ASD specific.

1.4.3.4 Fatima Talaat Autism Dataset

Fatima Talaat data set is among the few publicly accessible datasets that target individuals with Autism Spectrum Disorder (ASD). It describes six different facial expressions within the therapeutic context and indicates the subtle emotional responses that are sometimes uncharacteristic of non-autistic folks. Though small and scarce, it is a significant asset of emotion modelling specific to ASD [25]. It is not a generic dataset; it is ASD specific.

1.5 Problem Statement

Existing Facial Emotion Recognition (FER) systems are predominantly developed and evaluated using facial expressions of neurotypical individuals, with limited attention given to the distinct emotional expression patterns of individuals with Autism Spectrum Disorder (ASD). Consequently, these systems may not reliably

recognize emotions in ASD populations. This limitation can adversely affect the accuracy, robustness, and generalizability of FER systems for ASD population. Therefore, there is a need to investigate ASD-specific and diverse facial emotion data to achieve more accurate and reliable emotion recognition for individuals with ASD.

1.6 Research Objectives

The objectives of this research are as follows:

- i To improve ASD-focused FER accuracy by evaluating deep learning CNN models on ASD-specific and diverse facial emotion datasets.
- ii To improve FER generalizability for ASD populations using a hybrid dataset combining ASD-specific and general facial expression data.
- iii To improve FER performance by combining high-performing CNN models for more accurate and consistent emotion predictions.
- iv To compare individual and multimodel CNN performance using accuracy, precision, recall, and F1-score across ASD-specific, general, and hybrid datasets.

1.7 Research Questions

These questions will be answered through the implementation of the proposed system:

- i. How can deep learning-based CNN models improve facial emotion recognition accuracy for individuals with ASD?
- ii. To what extent does integrating ASD-specific and general facial emotion datasets improve the accuracy, reliability and generalizability of FER models?

- iii. To what extent can multimodel CNN fusion improve the accuracy and consistency of FER compared with individual CNN models?

1.8 Rationale and Significance of Selected Research Topic

Autistic people are also characterized by atypical facial expressions and a reduced ability to reciprocate emotions, which makes the use of traditional facial emotion recognition (FER) systems less effective with this population. The majority of the available FER models are trained on general datasets, including FER-2013 that fail to capture the individual emotional behaviours of an autistic person. This paper evaluates different deep learning models by using both general and autism-specific datasets to evaluate their performance. The outcomes are supposed to help develop more specific and individualized FER instruments to be used in clinical diagnostics and therapeutic settings.

1.9 Research Methodology

This study will follow a quantitative experimental approach focused on applying, evaluating, and comparing deep learning-based FER models for autism-aware emotion recognition using structured facial image datasets. Model performance will be assessed using standard evaluation metrics to ensure reliable and objective results.

The study will involve systematic data preprocessing, model training, validation, and testing under consistent experimental conditions. Different deep learning architectures will be compared based on their ability to accurately recognize and classify facial emotions. The findings will provide a quantitative basis for identifying the effectiveness and suitability of the selected models for autism-aware emotion recognition.

1.9.1 Problem Formulation

Identified the limitations of existing FER systems trained on general datasets and the need for improved emotion recognition models that account for the distinctive facial expressions of autistic individuals

1.9.2 Dataset Selection

Three dataset configurations were utilized:

- i. FER-2013: A widely used general-purpose facial emotion dataset.[\[26\]](#)
- ii. Autism Specific Dataset (Fatima Tallat / Kaggle ASD Dataset): A public dataset containing facial expressions of individuals with ASD.
- iii. Hybrid Dataset: A balanced dataset created by combining and harmonizing both general FER13 and autism specific ASD Fatima Tallat datasets to improve representation and reduce bias.

1.9.3 Data Preprocessing

All datasets were pre-processed using standard techniques, including image resizing, pixel normalisation, zero padding, imbalance checking, cleaning and categorical encoding of emotion labels, ensuring compatibility across datasets and models.

1.9.4 Model Selection and Architecture Design

Choose five CNN architectures known for their effectiveness in image classification:

- i. VGG16
- ii. ResNet50
- iii. DenseNet121

- iv. Xception
- v. MobileNet

1.9.5 Transfer Learning Implementation

It was employed by initialising models with ImageNet pre trained weights, followed by fine-tuning on the selected datasets to improve learning efficiency and performance.

1.9.6 Model Training and Validation

Each model was trained using consistent hyperparameter settings and validation strategies.

The datasets were systematically divided into training, validation, and test sets to ensure fair model comparison and reliable evaluation of generalization performance on unseen data.

1.9.7 Performance Evaluation

Performance was evaluated using standard quantitative metrics, including accuracy, precision, recall, and F1-score, to provide a comprehensive assessment of the models' classification performance.

1.9.8 Cross-Dataset and Cross-Model Comparison

The trained models were evaluated across all dataset configurations to:

Assess accuracy and generalization capability. Examine the impact of dataset composition, particularly the hybrid dataset, on recognizing subtle and atypical emotional expressions in autistic individuals

1.9.9 Hybrid Model Development

Based on the evaluation results, the best-performing architectures were identified, combined and used to design a hybrid model aimed at further improving recognition accuracy and reliability.

1.9.10 Outcome Analysis and Insights

The experimental findings were analysed to derive insights into optimal model dataset combinations and their applicability in ASD-focused FER systems. The methodology comprehensively addresses all stated research objectives and questions.

1.10 Important Findings

The key contributions of this study include:

- i Comparative analysis of the five deep learning architectures that are based on transfer learning to identify facial emotions, including VGG16, ResNet50, DenseNet121, Xception, and MobileNet.
- ii Augmented construction and analysis of a balanced hybrid FER dataset with general-purpose and autism-specific face expression data.
- iii Fatima Tallat data set that lacks validation-oriented research discussed.
- iv Introduced multimodel FER implementation and testing of the most performant architectures to enhance the recognition judgment.
- v A cross-dataset generalisation experiment of the strengths, weaknesses of general and autism-specific and hybrid datasets.
- vi Empirical recommendations to help inclusive, autism-wholesome FER systems to emerge in clinical, therapeutic, and education settings.

1.11 Summary and Thesis Outline

Chapter 1 gives the necessary background of the investigation, the problem statement, research objectives, summarises the methodology and the relevance of the facial emotion recognition (FER) to individuals with Autism Spectrum Disorder (ASD).

Chapter 2 has a detailed literature analysis, where research gaps in the existing FER techniques and datasets were identified, especially with regards to ASD.

Chapter 3 outlines the intended methodology, such as the type of datasets, model architectures, and the transfer learning mechanisms employed in the model of classifying emotions.

Chapter 4 outlines the implementation framework and details the processes of the experiment and outlines the tools and platforms that were used to train and evaluate the models also presents the outcomes of the experiment used to compare the performance of various models with the on general and autism-specific data exposing the results in detail and visualized examples. Chapter 5 wraps up the research, generalises the key findings, and suggests future research directions in establishing scaffolded FER system to support individuals with ASD.

Chapter 2

Literature Review

2.1 Introduction

This section reviews the major developments in Facial Emotion Recognition (FER), with a particular focus on its application to autism. The review is divided into three main areas. First, it discusses early studies that introduced and developed the concept of FER systems. Second, it examines recent research on FER for children with autism using machine learning and deep learning techniques. Finally, it provides a comparative analysis of existing approaches, focusing on model performance, datasets, and their relevance to autism. The section concludes by highlighting the key research gaps that can guide future work in autism-focused FER systems.

An important aspect of emotional intelligence is to identify feelings, which in turn is not only a biological / psychological process, but also a mental instinct to empathize and understand the emotional conditions of others. The ability to learn and correctly respond to facial expressions is a critical aspect of emotional intelligence, specifically, the ability to recognize the emotions of a face [27].

Over the years, scientists and techno gurus have diverted much of their energies towards developing facial detection systems that do not just simply find faces in photos.

These systems are increasingly skilled at examining the emotional state of a person through an evaluation of its facial features. The relative ease with which certain emotional expressions can be classified has led to development of many technologies using computers to enhance human computer interaction to enable machines to detect human attitudes and emotions and respond more precisely [28].

Over the recent past, studies in the examination of facial expressions in a person with autism spectrum disorder (ASD) have experienced growth significantly. Numerous researches have examined the challenges faced by children with ASD in terms of processing and deciphering facial expressions.

These issues may negatively affect their ability to interact in a social manner. Therefore, diverse intervention methods and technology solutions have been proposed, which would help children with disabilities (ASD) to better perceive and respond to different facial expressions.

2.2 Studies Introducing a Facial Emotion Recognition System in Neurotypical Individual

Automatic facial recognition was first explored by Bledsoe in the 1960s on behalf of the U.S. Department of Defence. The primary goal of this study was to determine the feasibility of using computational procedures in performing facial recognition operations [29].

Li et al. conducted a study on fair Facial Expression Recognition (FER). The purpose of the study was to reduce bias in FER systems because some facial expressions have fewer training samples than others. The objective was to develop a model that could recognize less-represented facial expressions more fairly. For the methodology, the researchers created a more balanced dataset using generated facial-expression images and developed a model called the Balanced Feature Fusion Network (BFFN). The model used facial action units and attention mechanisms to improve the recognition of difficult expressions. The researchers tested the

model on the RAF-DB and AffectNet datasets and found improvements in fairness and recognition performance. However, a limitation was that the study mainly focused on expression-class imbalance, while other types of bias, such as cultural and demographic bias, were not fully addressed. [30].

The Kanade formulated the first entirely automated system of face recognition whose purpose was to advance the emotions recognition capabilities of a machine by differentiating between the aspects of the face features that had been recognized by a machine with the face features as perceived by humans.

The process involved deriving sixteen unique face characteristics in order to decode emotional facial expressions. This system marked a vast step toward automatizing facial recognition; it was proved that a machine was able to trace facial expressions and other physiological changes and could effectively determine human emotions [31].

The initial concepts of Darwin stated that the mechanisms of emotion detection was complex and involved a number of models of behaviour which lead to a fine-grained classification of emotional conditions.

The objective of the experiment was to provide a scientific understanding of human feelings in relation to facial expressions through application of observation and relative analysis of facial expression in human beings and animals. Thus, the emotional classification system that Darwin formulated following his work included as many as 40 potential emotional states, a complete attempt at the states of emotions and a precursor of the later computational theory of emotion recognition [32].

Manalu and Rifai developed a hybrid CNN-RNN model for detecting human emotions from facial expressions in video data. The objective was to combine CNN-based facial feature extraction with RNN/LSTM-based temporal analysis. The researchers used the Emognition Wearable Dataset 2020 and classified nine emotions.

Three models were tested: custom CNN,RNN, MobileNetV2 RNN, and InceptionV3 RNN. The InceptionV3 RNN model achieved the highest testing accuracy

of 66%, followed by the custom CNN RNN at 63% and MobileNetV2 RNN at 59%. The main limitation was the relatively low recognition accuracy, particularly because the system had to distinguish nine different emotions.[33].

2.3 Studies Based on Emotion Recognition in Autistic Children

2.3.1 Using Machine learning approach

By analysing eye-tracking data, [34]investigated racial face processing differences in children with low-level autism. They used K-means clustering to select features and used a SVM classifier to classify between gaze behaviour of own-race and other-race faces. The model demonstrated high accuracy of 84.17% that indicates that the patterns of visual attention differ heavily and indicates the value of machine learning in the reporting of subtle patterns of social cognition dysfunctions in ASD.

In [35], author tested the hypothesis of micro-expressions as biomarkers to diagnose autism spectrum disorder (ASD). The fact that individuals with ASD usually experience an abnormal movement of facial muscles and the lack of expression of emotions reduced expressiveness and can be explained by the subtle micro-expressions became the guide to this research. The scholars sampled custom video to explore this. They used the common machine learning classifiers like Support Vector Machines (SVM) and Decision Trees to examine the temporal and spatial characteristics of these fleeting cues. The rationale of the methodology was rooted in the discovery of emotion-related micro-patterns which vary between neuro-typical and ASD subjects. Even though the study did not mention accuracy measures, the findings suggested that micro-expressions may be applied as non-invasive behavioural biomarkers of ASD.

The authors however made it clear that there was the need to experiment the reliability of these signals in clinical practise through larger and more controlled studies and multimodal integration.

A systematic review by [36] was carried out on 65 studies where machine learning was applied to recognize the feelings in the case of autistic individuals. The review suggested that methods of supervised machine learning were popular in previous FER research, such as SVMs, decision trees and naive Bayes. It indicated that the bulk of the research was on the six fundamental emotions, with the use of facial cues as the key input modality. According to the review, machine learning is a viable option in FER tasks, particularly when the size of the dataset or the computational capability can be a constraint on the use of deep learning.

In [37], Kiejdo, Depka Pradzinska, and Zawadzka investigated emotion recognition in children with Autism Spectrum Disorder using machine learning and multiple types of signals. The purpose was to improve emotion recognition by combining facial expressions with physiological information. The objective was to develop a multimodal system that could recognize emotional states more accurately than using facial expressions alone. For the methodology, the researchers combined facial-expression information with electrodermal activity, temperature, and heart-rate signals using a late-fusion machine-learning approach. The multimodal model achieved an accuracy of approximately 80%, showing that combining different signals can improve emotion recognition. A limitation is that the study used a relatively small participant group and required several sensors, which may make the system difficult to use in everyday environments.

2.3.2 Using a Deep learning approach

Traditional machine learning (ML) methods have several limitations compared with deep learning (DL) approaches, particularly in facial expression recognition (FER). ML algorithms often rely on manually engineered features, such as facial landmarks, texture, and geometric characteristics, which require domain knowledge and may fail to capture complex patterns in facial expressions. In addition, conventional ML methods have limited ability to learn hierarchical representations directly from raw images and may perform poorly when dealing with large-scale and unstructured visual data.

In contrast, DL models, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks, can automatically learn discriminative features and complex patterns from facial data. Therefore, although traditional ML methods may require fewer computational resources, their dependence on handcrafted features and limited representation learning capabilities can make them less effective for complex FER tasks.

Bhattacharjee et al. developed a hybrid deep-learning model to recognize emotions in autistic children while they interacted with a NAO humanoid robot. The purpose was to understand the emotional responses of children with autism during human-robot interaction.

The objective was to recognize subtle facial emotional cues using deep learning. For the methodology, the researchers collected around 50,000 facial frames from 15 autistic children and combined a ResNet-50 Convolutional Neural Network (CNN) with a Graph Convolutional Network (GCN) using facial landmarks. The model achieved an accuracy of approximately 90% for emotion recognition. However, the main limitation was the small number of participants, only 15 children, which limits how well the results can be generalized to the wider autistic population. [37].

In[38] Radočaj proposed an advanced deep learning approach for recognizing emotions in autistic children through facial expressions. The objective of the study was to compare CNN and Transformer-based architectures to determine which model provides more accurate emotion recognition for children with Autism Spectrum Disorder (ASD).

The research used a labeled dataset of facial images of autistic children representing emotions such as neutral, joy, anger, and fear. Among the evaluated models, the Swin Transformer achieved the best performance with approximately 80% accuracy. However, the study was limited by a relatively small and imbalanced dataset and relied only on facial images, which may reduce its effectiveness in real-world situations. The authors suggested using larger, more diverse, and multimodal datasets in future research to improve performance.

however, in [39] developed a real-time facial emotion recognition (FER) system that is specific to children with Autism Spectrum Disorder (ASD) to meet the requirement of having a specific and immediate emotion-detecting mechanism in the hands of therapeutical applications.

The experiment adopted a hybrid-automatizer, an autoencoder kernel of the convolutional neural network (CNN) where six main emotions, anger, fear, joy, neutral, sadness, and surprise, were employed. Xception model performed better than all the other deep learning models and yielded higher accuracy of 95.23 than MobileNet and ResNet, thus confirming its applicability in the real-time FER on autistic patients.

When recognising autism using the lastage of facial images, [40] conducted a comparative analysis to ascertain the effectiveness in deep learning models, VGG16 and MobileNet. The paper was meant to improve the accuracy of FER during ASD assessment procedures and utilized the sample of facial photographs to assess the operation of every model.

The results indicated that VGG16 was much better than MobileNet with when the accuracy (validation) was 99% and testing accuracy was 87%. The findings point to the utility of VGG16 to classify face-associated features in consideration of the emotional perception difficulties in autism.

Another study [41], “Assistive facial expression recognition for children with autism using re-enactment,” developed an assistive FER approach specifically for children with autism. The researchers combined facial re-enactment with a FER classifier to make expressions clearer. Their EfficientNet-v2 classifier achieved 67.83% accuracy for seven expressions and 64.65% for eight expressions. A clinical study included 48 participants, including individuals with ASD. Measuring the performance of deep neural networks in FER systems deployed in uncontrollable real-world conditions, [42] investigated the results of DenseNet and Xception. Their analysis aims at comparing the noise and obstruction error rate and the models robustness when operating in noisy and obstruction-prone environments such as those found in the natural environment.

DenseNet proved to be best at all in both accuracy and resiliency to distortions; Xception also demonstrated greater classification power than the base model. These findings indicate that FER can be used extensively in the two models in the non-controlled laboratory settings like those with autistic people.

In [38], had designed a mobile app, which was designed to make people with ASD identify facial emotions in real time. VGG-16 deep learning model and FER-2013 dataset have been added to the study to address the issue of poor engagement and emotional sensitivity of autistic users. The average F1 score of the application and its accuracy were 0.91. Indicative testing of three ASD participants showed positive outcomes as they demonstrated a higher level of interest and a higher level of communication, which proved the mobile tool to be effective in the process of recognition abilities of emotional situations in the real world.

A real time system proposed by [43] has worked out an emotion recognition system on children with autism. The issue which was discussed in this research was that the autistic children can also have the issues with recognition of emotions and their expression, it might negatively affect the social life of the child. The system involved three stages that included face recognition, feature isolation and a classification of these features. The system was trained on a deep convolutional neural network (DCNN) architecture to identify six basic emotions: anger, fear, joy, neutral, sadness and surprise.

The findings showed that the DCNN-based model is effective in the context of emotional expression classification that may be implemented as a valuable support mechanism to improve emotional cognitions in children with autism.

In other places, [44] suggested a multimodal system [AEGIS] in real-time and the system employs augmented reality or computer vision to assist in detecting facial emotions in individuals with ASD. They performed an investigation that focused on the challenges facing majority of autistic individuals to decode the emotional messages. The solution consisted of computer vision with the DCNNs to recognise the facial expressions and analyze them in real-time, and the augmented reality gave real-time feedback to the user.

Those results suggested that AEGIS increased the degree of emotional responses and accuracy in recognition, which is why the tool can be described as a promising one that can assist individuals with ASD imagine their feelings better.

Akalya in [45] conducted a study on the use of deep learning for Autism Spectrum Disorder (ASD) detection and emotion recognition in children. The purpose of the study was to develop an AI-based method that could identify autism-related patterns and recognize emotions from facial images. The objective was to improve the performance of deep-learning models for both ASD detection and facial emotion recognition. For the methodology, the researchers used two public datasets and applied preprocessing techniques such as normalization and data augmentation. They developed a hybrid ResNet50V2 + InceptionV3 model for ASD prediction and an enhanced MobileNet model for recognizing six emotions in children with ASD. The hybrid model achieved 95% accuracy for ASD prediction, while the enhanced MobileNet achieved 73.3% accuracy for emotion recognition. The study showed that deep learning can provide a non-invasive approach for facial-image analysis in autism. However, a limitation is that the models were trained using public datasets, so further testing on larger, autism-specific and real-world datasets is needed before clinical use.

Monani et al. [13] evaluated deep-learning models for autism detection using children's facial images. The purpose was to determine which CNN architecture performs best for ASD classification.

The researchers used a balanced dataset of 5,000 facial images, consisting of 2,500 ASD and 2,500 non-ASD images. They compared VGG16, VGG19, ResNet50, InceptionV3, and MobileNet, using transfer learning and data augmentation.

The models were evaluated using standard classification metrics to measure their effectiveness and reliability in distinguishing between ASD and non-ASD facial images. The study demonstrated the potential of deep learning and transfer learning techniques for automated autism detection. Their findings provide a valuable foundation for further research and development of computer vision-based approaches for autism-related analysis, particularly in exploring more accurate and effective automated methods.

2.4 Comparison of the Articles

The core approaches, methodologies, technologies, and techniques applied in the above-mentioned articles have been systematically classified and presented in tabular form to provide a clear and structured overview of the existing research. This classification enables the reviewed studies to be easily distinguished and compared based on their respective research objectives, methodologies, datasets, models, and technologies.

The purpose of organizing the previous studies in this manner is to provide a comprehensive record of the existing work and to examine how different researchers have addressed similar research issues and challenges. By comparing these studies, the similarities and differences in their proposed approaches, experimental procedures, and reported outcomes can be identified more effectively. [2.1](#).

2.5 Identified Research Gap

from different other studies like [\[46\]](#), [\[47\]](#), [\[48\]](#), [\[49\]](#), [\[50\]](#), [\[51\]](#) [\[52\]](#), [\[53\]](#), [\[26\]](#), [\[54\]](#), we can notice, Existing Facial Emotion Recognition (FER) systems have primarily been developed using generic facial-expression datasets, which may not adequately capture the distinctive facial characteristics and emotion-expression patterns observed in individuals with Autism Spectrum Disorder (ASD).

Although some studies have developed FER systems using ASD-specific datasets, these datasets are generally small, limited in diversity, and collected under constrained conditions. Consequently, models trained exclusively on such datasets may suffer from overfitting and have limited generalization to unseen individuals and real-world environments.

Therefore, a significant research gap remains in developing a reliable, accurate, and generalizable FER system for individuals with ASD that can address the limitations of both generic and small ASD-specific datasets.

In particular, there is a need for approaches that can effectively leverage available

Table 2.1: Summary of Related Work on Facial Emotion Recognition for Autism Spectrum Disorder

Ref.	Author(s)	Year	Dataset	Model / Method	Accuracy	Research Gap
[26]	S. K. S. et al.	2021	Not specified	Deep CNN (three-stage FER: face detection, feature extraction, emotion classification)	Not specified	No benchmarking, lacks validation metrics and dataset transparency.
[27]	Kondaveeti & Goud	2022	CK+, JAFFE	VGG16, ResNet152V2, InceptionV3, Xception	93.62%, 91.25%, 89.14%, 92.34%	General FER only; not autism-specific.
[28]	F. M. Talaat	2023	ASD Kaggle Dataset	Xception, MobileNet, ResNet with Kernel Autoencoder	95.2%, 88.12%, 91.43%	No comparison using multiple benchmark datasets.
[43]	Canedo & Neves	2023	AffectNet, RAF-DB	DenseNet, Xception	~93%, ~91%	No evaluation on ASD population.
[35]	Rizka Wakhidatus et al.	2024	FER-2013	VGG16	91.0%	Needs deeper focus on autism-specific behavioural analysis.
[41]	Noura Alhakbani	2024	FER-2013, Custom ASD Dataset	ResNet50, VGG16	92.3%	Limited cross-dataset performance analysis.
[39]	Hu et al.	2024	Custom Video Samples	SVM, Decision Tree (Micro-expression Analysis)	Not specified	Exploratory study requiring larger clinical trials.
[38]	Petra Radočaj and Goran Martinović	2025	FERAC Dataset (770 images, Kaggle)	InceptionV3, InceptionResNetV2, DenseNet201, ResNet152V2, Swin Transformer	74.84%, 73.55%, 72.90%, 71.61%, 80.00%	Results are limited by a small and imbalanced dataset.
[44]	Nilay Kale	2025	Custom ASD Facial Dataset (derived from multiple sources)	VGG16	79.0%	Dataset sources are not fully specified, limiting reproducibility.
[46]	Akalya et al.	2026	Public facial-image dataset from Kaggle	ResNet50V2 + InceptionV3,enhanced MobileNet	73.3%	Dataset sources are not fully specified..

ASD-specific information while reducing overfitting and improving the model's ability to generalize across individuals, facial variations, and real-world conditions.

Table 2.2 offers an effective summary of the latest research in facial emotion recognition (FER), highlighting their use of popular deep learning models, such as VGG-16, ResNet, MobileNet, Xception, and DenseNet, the types of data they used, their consideration of autism, and the presence of emotion recognition.

It also provides a clear basis for comparing existing approaches and identifying gaps for future research.

Specifically, the study by Dr. Fatma M. Talaat that uses the Fatima Talaat dataset

applies the Xception and MobileNet architectures using facial data that is autism-specific, in line with the aims of emotion recognition. Contrarily, other authors such as Sohail Masood and co. explore a broader range of models, yet they do not consider the emotional classification.

Table 2.2: Most Relevant Studies from Table 2.1 Summary

Authors	Year	Dataset	VGG 16	Res Net	Mob Net	Xcep tion	Dense Net	Autism Spec.
Noura Alhakbani	2024	FER-2013, Custom ASD	✓	✓	✗	✗	✗	✓
Dr. Fatma M. Talaat	2023	ASD dataset	✗	✗	✓	✓	✗	✓
Rizka Wakhidatus	2023	FER-2013	✓	✗	✗	✗	✗	✗

2.6 Summary

In this chapter, all the papers that were closely associated with the proposed architectures and applicable techniques that have been studied in the area of study were reviewed. The next chapter will provide a summary of the suggested methodology and other technologies that will be used to evaluate the results based on the underlying article.

Chapter 3

Methodology

The following section will address the methods applied in the proposed system. The main objective is to test the validity of five different deep learning models on the particular autism, generic, and hybrid datasets; all the steps that arrived at the results are going to be described.

3.1 Introduction

This research has sought to enhance the automatic recognition of emotions based on facial images in the previous chapters, particularly amongst people with Autism Spectrum Disorder (ASD). The reason why many current facial emotion recognition systems find it difficult to recognize autistic facial expressions is due to the fact that most of them are usually trained on ordinary datasets. To this end, various well-established deep learning architectures including VGG16, ResNet50, DenseNet121, MobileNet, and Xception were studied and trained with the help of transfer learning. These models were applied to three types of data: general facial emotion dataset (FER-2013), autism-specific public dataset, as well as an original hybrid dataset that brings the two together. The system also managed to learn significant facial features using pre-trained ImageNet models without the need to have incredibly large training set.

Special emotion recognition classification layers were introduced to each of the models. Accuracy, precision, recall and F1-score were used to compare model performance, paying special attention to how each method performs with respect to identifying subtle, atypical emotional displays that are typically common in autistic people. A hybrid model was derived after determining the most effective architectures in order to enhance accuracy of recognition even more. On the whole, the results demonstrate that using autism specific and general data on faces, combined with transfer learning and hybrid modelling, results in more valid and inclusive facial emotion recognition systems.

3.2 Objective

The main aim of the current research as shown in Figure 3.4 is to explore the impact of a dataset selection and quality on the level of facial emotion recognition performance among sensitive population, specifically those with Autism Spectrum Disorder (ASD), and to determine the best deep learning models and dataset configurations to get good and valid results.

3.3 Model Selection

Purposed model was trained with five prominent deep learning models: VGG16, ResNet, DenseNet, MobileNet, and Xception. Each model utilized a transfer learning approach, starting with their pre-trained versions from the extensive ImageNet dataset. This approach provided a solid groundwork for the models by allowing them to grasp fundamental patterns and features (such as edges and textures) without needing to start from the beginning.

while the initial layers were left unchanged in both models, as these components are already adept at detecting general visual features.

After that modified the final layers to focus on classifying images into the seven emotion categories present in my database. In certain cases, then fine-tuned the

deeper layers of the model to enhance performance, particularly when the pre-trained features did not fully correspond with the emotion detection task. Each model brought distinct advantages:

VGG16 features a straightforward and deep architecture, making it easy to adapt.

ResNet incorporates shortcut connections that simplify the training of deeper.

DenseNet connects all layers directly, which enhances feature reuse.

MobileNet is a compact and efficient model, advantageous in scenarios with limited hardware.

Xception employs advanced convolutions to improve both speed and accuracy. Recent relevant research claims that these selected models achieve the highest accuracy.

Experimenting with different models allowed me to evaluate their strengths and identify which one performed optimally for emotion classification in this instance [55].

3.4 Model Architecture

Each of the five models—VGG16, ResNet, DenseNet, MobileNet, and Xception was adapted using a similar modification strategy to suit the facial emotion classification task.

The pre-trained convolutional layers of each model were utilized to capture essential visual features. For the transfer learning process, the lower layers were either locked or partially fine-tuned based on the training results and the specific model.

Additional fully connected (dense) layers were added on top of each pretrained base to recognize patterns pertinent to the task. To mitigate overfitting, a dropout layer was sometimes incorporated, which randomly deactivates some neurons during training.

Every model ends with a softmax output layer, which translates the final predictions into one of the seven emotion categories: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral.

Despite their architectural variations, like ResNet’s skip connections, DenseNet dense connections, MobileNet’s lightweight structure, and Xception’s efficient convolutions, all models followed this general framework to ensure consistency in testing.

3.5 Model Training

3.5.1 Employing Pretrained Architectures

The models of different major convolutional neural network (CNN) constructions were trained to classify emotions based on a facial image: VGG16, ResNet, DenseNet, MobileNet and Xception. These models were trained using already pretrained weights based on the ImageNet dataset, which is why they can be used as functional feature extractors. The only modification that was made to the architecture was in the upper levels of the classification, in emotion detecting.

3.5.2 Application of Transfer Learning and Feature Extraction

The transfer learning technique was applied, in which the convolutional base of all the already trained models is frozen during the training. This meant that the mass of these layers had not been changed so that the models did not forget the generalised visual attributes that they had been trained on in large natural image collections.

The only learning that was done involved training the recently added custom classification head and it was possible to extract task specific features applicable in emotion classification.

This method has the advantage of being applicable in situations where only a

small amount of data is available since it enhances generalisation and minimises chances of overfitting.

CNN-based model was employed to automatically extract high-level features from the input images. Unlike traditional feature extraction, CNNs learn relevant features directly from the image data through convolutional layers, reducing the need for manual feature engineering.

3.5.3 Custom Classification Head

A traditional head of classification was introduced to tune the pretrained models. This usually involved:

A Global Average Pooling (GAP) layer to downsample the size of the feature maps, multiple Dense (fully connected) layers to learn at least some high-level abstractions,

A final Dense layer with Softmax activation to generate probabilities in the seven categories of emotions this person is heating up about (Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral).

3.5.4 Model Compilation

All models were set with the following parameters:

- i. Loss Function: Categorical Cross-Entropy helps minimize the difference between predicted and actual class probabilities, enabling the model to learn the correct class for each input more effectively.
- ii. Optimiser: Adam was chosen due to its efficiency and capability to adaptively change the learning rate during training, helping the model converge faster and achieve stable performance.
- iii. Evaluation Metric: Accuracy measures the model's performance during training and validation by determining the proportion of correctly classified samples out of the total samples.

3.5.5 Training Procedure

The number of epochs trained on the models was fixed at 10-20 with a batch size of 32. The training data were further complemented with validation data that would be utilized to monitor the model performance when applied to actual application and the problems that might be related to overfitting. Indicators such as loss and accuracy of the training were captured at each epoch.

3.5.6 Model Saving

After the training had been finished, all the models were saved in the h5 format to Google drive to be used in future or revisited. Despite the more recent versions of TensorFlow becoming compatible with the .keras format, the most widely used and universally used format is the .h5 format. It preserves the model architecture and learned weights, enabling easy reuse, deployment, and future experimentation.

3.6 Model Training on Individual Dataset

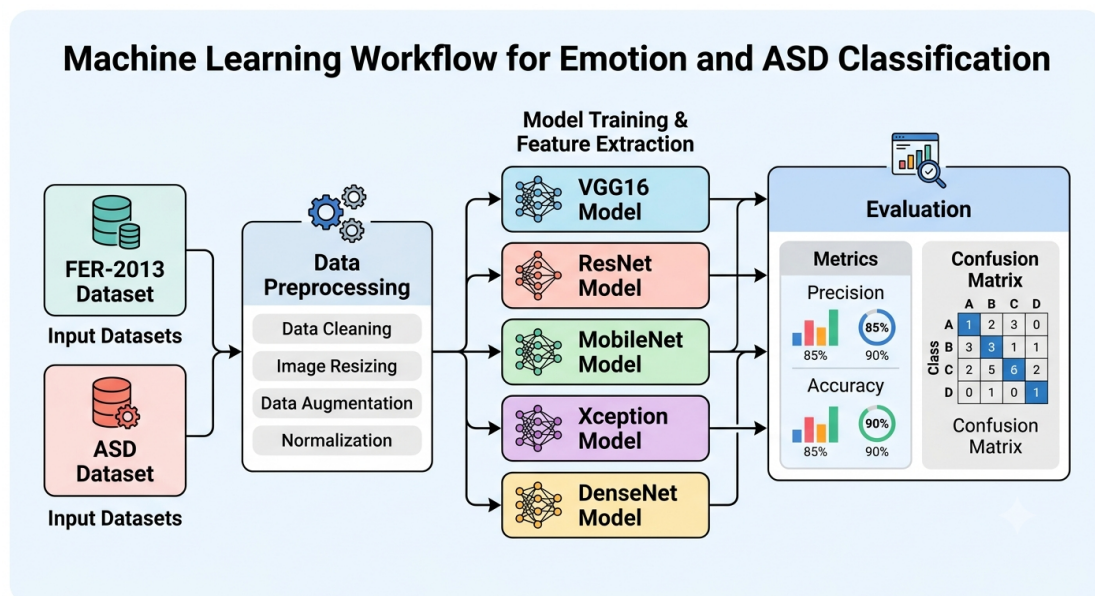


Figure 3.1: Model training on FER13 and ASD Fatima Tallat, individual datasets

As shown in fig 3.2 We have selected two raw datasets, one that focuses specifically on the emotions of autistic children, which is publicly available and has been recently utilized in relevant ASD studies, known as FTD. Additionally, we have chosen a benchmark dataset that is well-established, publicly accessible, and applied in FER systems, although it is not directly related to ASD emotions, referred to as FER2013.

3.7 FER13

This dataset consists of grayscale facial images with a resolution of 48×48 pixels. Each face has undergone preprocessing and alignment to appear approximately centered and uniformly scaled across all images. The aim is to classify each image based on the emotional expression it conveys, categorizing it into one of seven emotion types as Angry , Disgust , Fear , Happy , Sad , Surprise , and Neutral . The dataset is split into two parts, a training subset with 28709 samples and a public test subset containing 3589 samples.

3.8 Fatima Tallat’s Dataset for Autistic Children

This dataset contains six folders, each representing a distinct class of emotions, including natural, sadness, fear, joy, anger, and surprise, totaling 1400 images divided into training and testing sets.

3.8.1 Dataset Preprocessing

Each dataset underwent the same preprocessing pipeline to ensure consistency throughout the study. First, all facial images were resized to 48×48 pixels, providing a uniform input size while reducing computational complexity. The pixel values were then normalized to a range of 0 to 1, which improved training stability and accelerated model convergence.

To enhance the diversity of the training data and reduce the risk of overfitting, several data augmentation techniques were applied, including random rotations, horizontal flipping, width shifts, and height shifts. These transformations artificially increased the variability of the training samples while preserving the underlying facial expressions.

Following preprocessing, each dataset was divided into training and testing sets using an 80:20 ratio. From the training set, a portion of the data was further reserved for validation during model training. This splitting strategy ensured that sufficient data were available for learning while maintaining independent datasets for validation and performance evaluation.

The same preprocessing, augmentation, and data-splitting procedures were consistently applied across all datasets to ensure fairness and uniformity. Consequently, any observed differences in model performance can be attributed to the proposed methodology rather than inconsistencies in data preparation.

3.9 Noise Removal and Filtering:

Started by checking how well the model did on each class to spot any confusing or tricky samples . When we found images that were mislabeled, misleading, corrupted, or just plain noisy, we took them out. That way, the dataset stayed clean and solid .To cut down on even more noise, we used smoothing filters, resized the images, normalized them, and added zero padding. This kept all the input sizes the same without messing up the images. Here also did data augmentation rotating, flipping, and tweaking the brightness. That gave more variety, helped prevent overfitting, and pushed the model’s accuracy higher.

Although several automated image denoising and quality assessment techniques, such as Gaussian filtering, Median filtering, Non-Local Means, and blur detection, are available for removing noisy or low-quality images, they were not used in this study. Facial emotion recognition relies on subtle facial features, and applying such techniques may alter or smooth important expression details, potentially

affecting model performance. Since the dataset size was manageable, all images were manually inspected, and those with excessive noise, severe blur, occlusions, poor face visibility, or incorrect labels were removed. Manual screening ensured the retention of high-quality images while preserving the original facial expressions, resulting in a more reliable dataset for training and evaluation.

3.10 Model Training on Hybrid Dataset

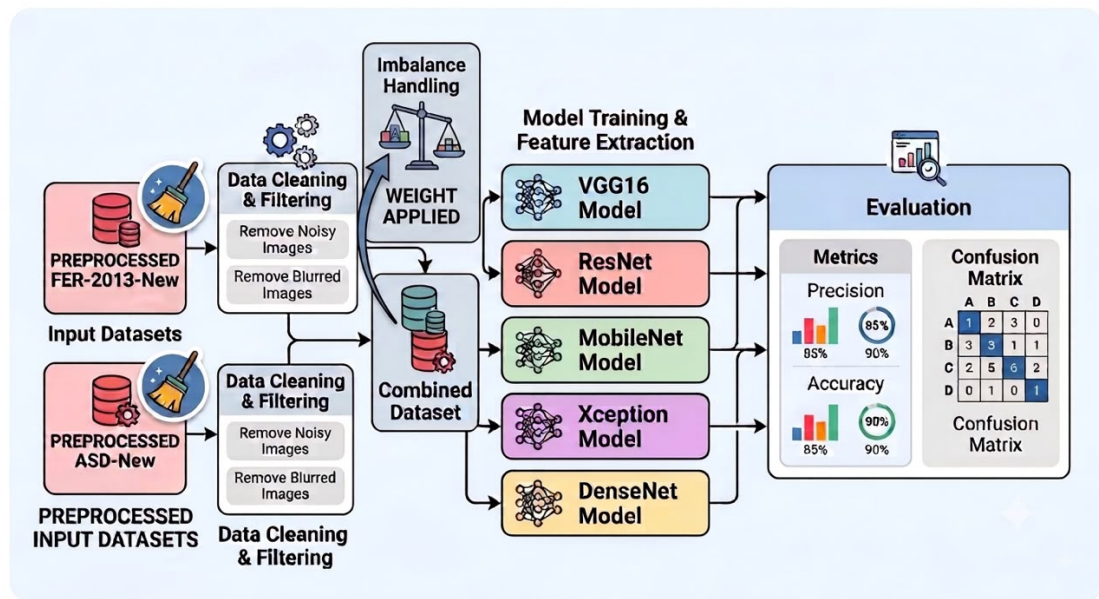


Figure 3.2: Model training mechanism for proposed system using hybrid dataset.

This study can not rely on ASD dataset only, the dataset's small for teaching a model to spot facial expressions. When you don't have enough data, the model tends to memorize the training faces instead of actually learning how to handle new ones. This is a real headache in facial expression recognition, since people's faces change so much with different lighting, angles, and just being themselves.

There's another problem, too. Most of the autism-specific data only shows expressions that are common in children with ASD. So, if a child with ASD happens to show a more typical expression, the model might not catch it. That means it won't work as well for everyone, and its accuracy takes a hit. This limitation can make the model less reliable when dealing with diverse emotional expressions and

individual differences among children with ASD. Therefore, using more diverse and representative datasets is important for improving the generalizability and overall performance of autism-aware emotion recognition models

3.10.1 Hybrid Dataset

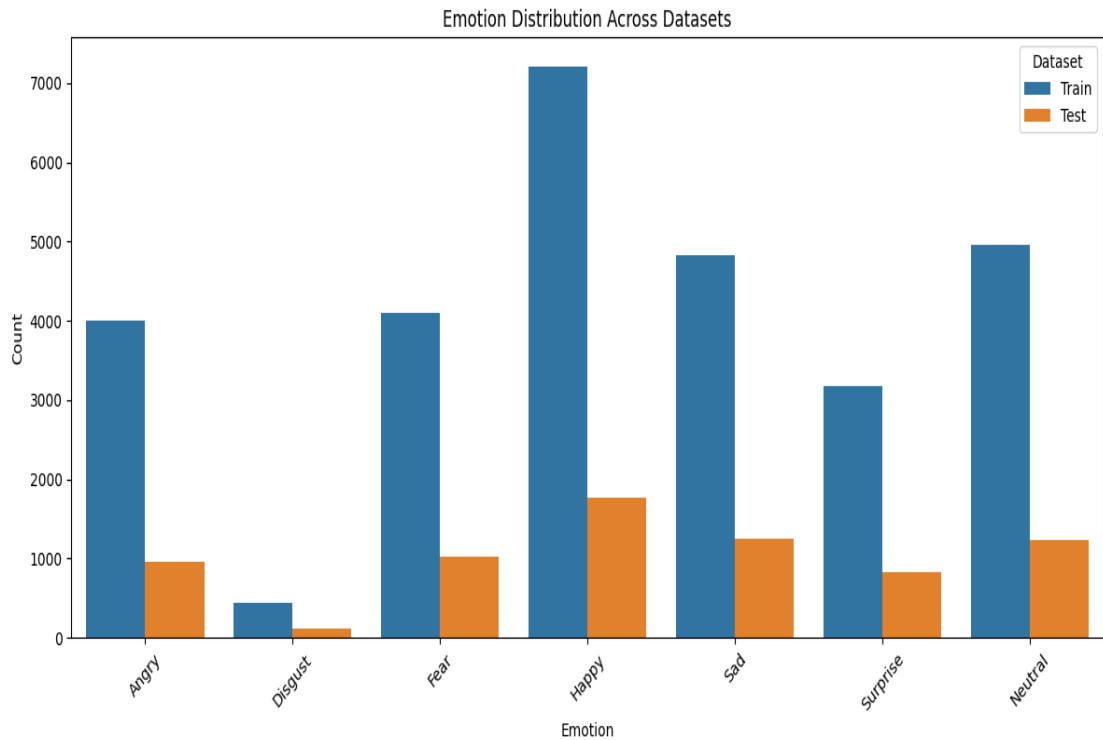


Figure 3.3: Bar Chat for Emotion Distribution Across Hybrid Dataset showing class inbalance.

A new hybrid dataset was developed by merging both ASD specific and generic datasets. Before integration, systematic data cleaning was performed to remove low-quality samples, such as blurry, noisy, ambiguous, or poorly oriented images, ensuring that only high-quality and representative data were retained. 28,709 training images from FER 1397 samples from ASD Fatima tallat dataset .This refinement process resulted in a cleaner and more reliable dataset suitable for model training and evaluation. The finalized dataset remained class-imbalanced, with happy being the most represented class (25.05%) and disgust being the least represented (1.52%). Visual inspection As shown in was conducted to further

analyze dataset characteristics and to present sample images for each emotion category, highlighting the existing imbalance like in figure 3.3 .

Following dataset construction, several preprocessing operations were applied. Pixel values were normalized to the range $[0, 1]$, and emotion labels were converted into one-hot encoded vectors for multi-class classification. The dataset was then shuffled to minimize ordering bias and partitioned into training and testing subsets using an 60/20/20 split ensure reliable model development and evaluation. The training set (60%) was used to learn the model parameters, the validation set (20 %) was used to tune hyperparameters, monitor performance during training, and implement techniques such as early stopping, while the test set (20%) was reserved for the final unbiased evaluation of the model. This three-way split helps prevent overfitting by ensuring that model tuning is performed on the validation set rather than the test set, making the final test accuracy a more reliable measure of the model's generalization performance.

3.10.2 Handling Class Imbalance for Hybrid Dataset

To handle class imbalance, common approaches like oversampling, under sampling, and synthetic data generation methods such as SMOTE or ADASYN each have drawbacks. Oversampling can lead to overfitting, as the model may memorize duplicated minority samples rather than generalize. Under sampling risks information loss by discarding valuable majority class data, potentially reducing model performance.[45]

SMOTE and ADASYN generate synthetic samples, but these can sometimes be noisy or unrealistic, which may confuse the model.[13][56]

To avoid these issues, after that the class weight technique was choosen, which adjusts the model's learning process to pay more attention to minority classes without altering the dataset, thus preserving all data while addressing imbalance effectively.To handle class imbalance, I used class weights, calculated with the formula:

$$W_c = \frac{N}{K \cdot N_c}$$

where N is the total number of samples, c is the number of classes, and k is the number of samples in class. This gives higher weight to minority classes, helping the model focus on them without changing the original data.[30]

3.11 Multi-Model Training on Hybrid Dataset

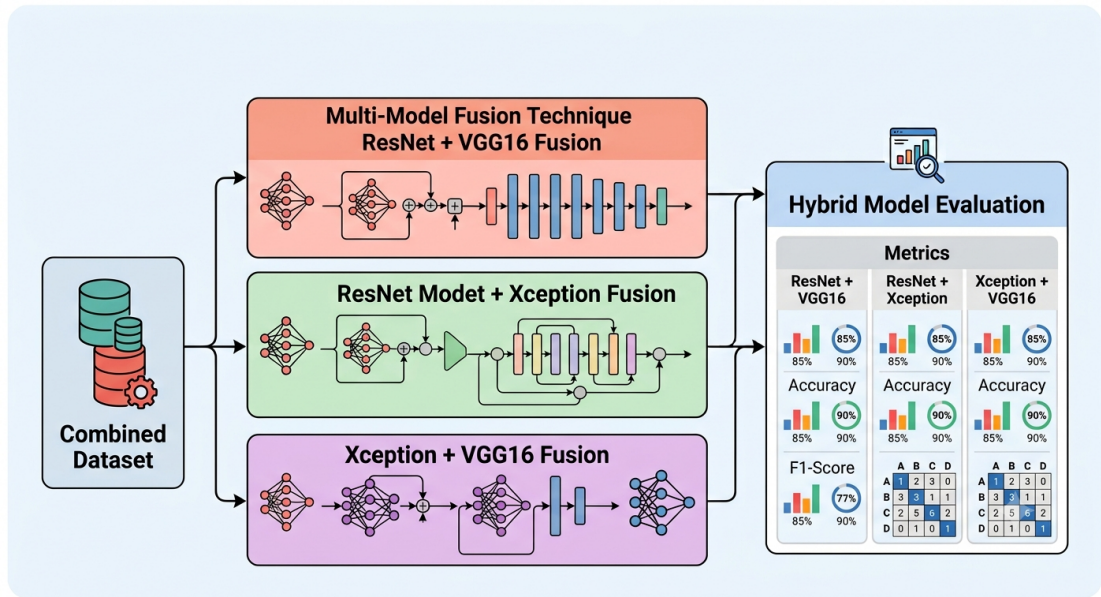


Figure 3.4: Multi-model based training mechanism on hybrid dataset

A multi-model approach shown in figure 3.4 for facial emotion recognition (FER) leverages the unique strengths of multiple models to achieve more accurate and robust results. **VGG16** captures fine-grained facial textures such as eyes, mouth corners, and subtle muscle movements, which are crucial for distinguishing subtle emotions. **ResNet**, with its residual connections, excels at learning deeper and more abstract patterns, making the system robust to variations in pose, lighting, and facial orientation. **Xception**, which uses depthwise separable convolutions, is highly efficient and effective at capturing complex spatial correlations in facial features while reducing computational overhead compared to traditional CNNs.

By combining these models in a hybrid architecture, the system benefits from **richer feature representations** that encompass both shallow, detailed textures and deep, abstract patterns. This multi-perspective approach improves overall **accuracy**, particularly for subtle or ambiguous emotions, enhances **robustness** to dataset variations such as occlusion or lighting changes, reduces **overfitting** on small datasets, and allows flexible **transfer learning** by leveraging pre-trained weights from each network.

In essence, hybrid model pairs combine the complementary strengths of different neural network architectures, allowing them to compensate for each other's weaknesses. By integrating diverse feature extraction and learning capabilities, these hybrid systems can capture a wider range of facial expression nuances.

This collaboration enhances the overall robustness of the facial emotion recognition (FER) system, making it more adaptable to varied datasets and real-world scenarios. As a result, hybrid models tend to achieve higher accuracy, improved generalization, and greater reliability compared to individual models operating in isolation. This synergistic approach ultimately leads to more consistent and effective emotion detection performance across different populations and conditions.

3.12 Multi-model Model Architecture

In figure 3.5 This hybrid CNN architecture mixes parallel and sequential paths to pull out features from images. First, The model begins with a Conv2D Convolutional layer, which is used to automatically extract important features from the input image, such as edges, corners, textures, and shapes.

Unlike traditional methods that require manual feature extraction, Conv2D learns these features automatically during training, making it highly effective for image classification tasks and other computer vision applications.

After convolution, Batch Normalization (BN) is applied to normalize the output of the previous layer. This stabilizes the learning process, speeds up model convergence, and reduces the chances of vanishing or exploding gradients.

It also provides a slight regularization effect, helping the model generalize better.

The output is then passed through a MaxPooling2D layer, which reduces the spatial dimensions of the feature maps by keeping only the most important features. This decreases computational cost, reduces memory usage, and makes the model more robust to small shifts or distortions in the input image.

The architecture then uses parallel feature extraction paths to capture information at multiple scales. The lower-resolution path focuses on broad contextual patterns and global structures, while the higher-resolution skip path preserves fine details that might otherwise be lost during pooling.

This multi-scale design improves the model's ability to recognize objects with varying sizes and levels of detail.

The features from both paths are merged using Concatenation, which combines information from different resolutions into a single feature representation.

This allows the network to use both detailed and contextual information simultaneously, resulting in richer and more discriminative features.

The stride-based Conv2D (Mixer) in the hybrid reduction stage is used to further extract features while reducing the spatial dimensions of the feature maps.

This layer efficiently mixes information from different scales and prepares compact, informative representations for deeper processing.

A skip connection is included to pass important features directly to later layers without modification. Skip connections help preserve useful information, improve gradient flow during training, and reduce the risk of information loss in deep networks.

Before classification, the feature maps are converted into a one-dimensional vector using the Flatten layer. Flatten does not learn any parameters; it simply reshapes the extracted features into a format that can be processed by fully connected layers. This allows the important spatial features extracted by the convolutional layers to be passed efficiently to the classification stage. The flattened feature vector is then used by the dense layers to predict the final output class.

The flattened features are then passed through a Dense (Fully Connected) layer with ReLU activation. This layer learns high-level relationships between the extracted features and the target classes, while the ReLU activation introduces non-linearity, enabling the model to learn complex decision boundaries.

A Dropout layer is added after the dense layer to reduce overfitting. During training, Dropout randomly deactivates a fraction of neurons (e.g., 50 percent), preventing the network from relying too heavily on specific neurons and improving its ability to generalize to unseen data. Finally, the model uses a Softmax output layer with 7 neurons, where each neuron corresponds to one class. The Softmax activation converts the outputs into probability values that sum to one, and the class with the highest probability is selected as the final prediction.

3.12.1 Frozen Layers

The initial Conv2D layers of the pre-trained backbone (along with their associated Batch Normalization layers, if applicable) are kept frozen during the initial training stage. This helps preserve the useful low-level features learned from the original training dataset while reducing the risk of overfitting. The remaining layers are then trained to adapt the model to the specific requirements of the target emotion recognition task. These layers have already learned useful generic visual features, such as edges, textures, shapes, and basic patterns, from large-scale datasets like ImageNet. Since these low-level features are transferable to many different image classification tasks, freezing the layers allows the model to reuse this previously learned knowledge without modifying their weights. This also reduces the number of trainable parameters, making the training process faster and helping to reduce the risk of overfitting. Freezing the early layers reduces the number of trainable parameters, leading to faster training, lower computational cost, and a reduced risk of overfitting, especially when the target dataset is relatively small. It also preserves the robust low-level features learned during pre-training while allowing the newly added layers to learn task-specific features for the seven-class classification problem.

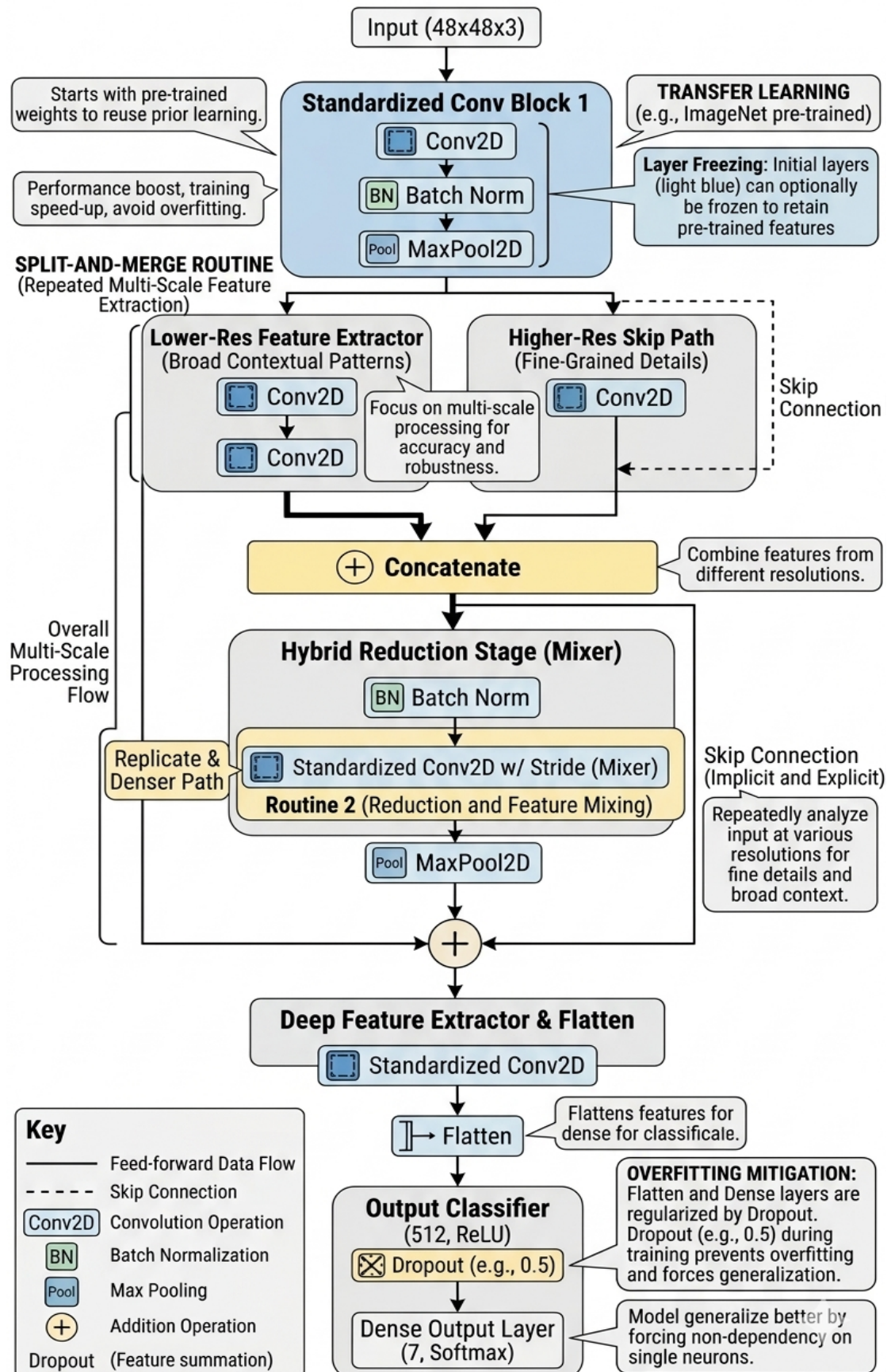


Figure 3.5: Multi-Model CNN based Architecture

3.13 Performance Metrics

To evaluate the effectiveness of the models of emotion classification to the full extent, several commonly used classification measures were used:

- i. TP (True Positives): Samples correctly identified as belonging to the respective category.
- ii. FP (False Positives): The samples that were falsely determined to be a part of the class.
- iii. FN (False Negatives): Actual data of the class that were incorrectly predicted.
- iv. TN (True Negatives): The samples were correctly recognised as not being members of the class.

3.13.1 Accuracy

The proportion of the labels of emotions that were correctly predicted to the overall number of predictions.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (3.1)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.2)$$

3.13.2 Precision

The rate of the correct predictions of the instances of a certain class out of all the predictions of the model concerning that class is the measure of the precision of this model.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3.3)$$

3.13.3 Recall (also referred to as Sensitivity)

The number of the correct predictions of instances of a particular class compared with the number of actual instances of that class, which measures whether the model can identify true positives.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3.4)$$

3.13.4 F1-Score

The harmonic combination of precision and recall that balances the two and is very useful in datasets with uneven distribution of classes.

$$\text{F1-Score (Formula 5)} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.5)$$

These measures will provide significant information about the total ability of the model, as well as the recognition ability of this or that emotion category. Macro F1 is used to measure overall performance by counting all classes' TP, FP, and FN together. It is useful when you care about overall prediction performance, and in single-label multiclass classification, Micro F1 = Accuracy. Weighted F1 is used when classes are imbalanced, so larger classes contribute more to the overall score.

3.13.5 Confusion Matrix

Each model was put in a confusion matrix to indicate its performance to classify all the seven categories of emotions. The matrix provides a holistic picture of how each category is actually doing in terms of its true positive, false positive and false

negative and allows an understanding of which categories that the model is failing to perform well on or misreading related emotions.

Table 3.1: Confusion matrix showing placement for samples

	Predicted: Happy	Predicted: Sad
Actual: Happy	TP	FN
Actual: Sad	FP	TN

3.13.6 A Training and Validation Curve

A training/validation learning curve is a graph that shows model performance (such as accuracy or loss) on the training and validation datasets across epochs to monitor learning and detect overfitting..

3.14 Tools and Environment

The study implementation and experimentation were performed using Python programming language, primarily utilization of the TensorFlow and Kerasdeep learning libraries defining and training convolutional neural networks. The experiments were run on Google Colab, which is a cloud computing development environment, with embedded support of GPUs. In particular, Tesla T4 (GPU) was made use of during the training process to speed up the process of calculations, a component of deep learning models, to a significant extent.

Alongside TensorFlow and Keras, a variety of other Python libraries were used, including:

- i. NumPy for carrying out numerical operations.
- ii. Matplotlib and Seaborn for visualising data.
- iii. Pandas for data manipulation and analysis.
- iv. Scikit-learn for evaluation metrics and creating confusion matrices.

- v. OpenCV for preprocessing pipelines , image resizing and face detection

This mix of tools and environment created a strong and efficient platform for developing, training, and evaluating deep learning models focused on facial emotion recognition.

3.15 Summary

The presented methodology explaining all the elements and instructions of the TC unit was applied in practice when the schedulers were employed in the experiments. The following chapters will provide the codes of the commands used and the utilised topologies used in the simulations. The objective is to obtain conclusions that will support the hypothesis.

Chapter 4

Results and Discussions

In this chapter, the experimental implementation and the results and discussion of the deep learning approaches put forward in this study are provided.

Although the last chapter has detailed the model architectures, datasets, and evaluation strategy, this chapter dwells on the real-world performance of the given models in the context of facial emotion recognition (FER) and especially to individuals with Autism Spectrum Disorder (ASD).

VGG16, ResNet50, DenseNet121, MobileNetV2, and Xception are five transfer learning-based CNN models that were implemented and tested using three dataset setups, namely the general FER-2013 dataset, an autism-specific public dataset (Fatima Tallat ASD), and a balanced hybrid dataset consisting of the two.

Based on these comparative results, a hybrid FER model was developed by combining the most effective architectures identified during the evaluation. The proposed hybrid model was then trained and tested on the hybrid dataset to determine whether the combination of different architectural strengths could improve overall facial emotion recognition performance. Accuracy, precision, recall, F1-score, confusion matrices and per-class performance were used to examine model performance. The findings show the efficacies of both the integration of data and hybrid modelling, and they are also effective to enhance robustness and inclusivity to ASD-conscious FER systems.

4.1 Implementation Environment

The experiment applied the experimental implementation to a controlled and consistent setting to encourage equal evaluation of all deep learning models. Google Colab Pro+ was selected as the most popular platform to develop on due to its ease of access, scalability and compatibility with a graphics card to speed up its processing. All training and testing experiments were executed in a Tesla T4 with 16 GB of dedicated memory and 16GB of system RAM. This setup had sufficient computation power to accommodate the large FER-2013 data set and hybrid dataset, and allow shorter training periods and fine-tuning of the models.

This was also accompanied by the popularity of the machine learning and computer vision libraries that were part of the software stack. The major concepts applied to build, train, and refine the CNN structures were TensorFlow and Keras. Evaluation metrics, confusion matrices and classification reports were also calculated using scikit-learn. Visualisations including accuracy/loss curves and performance per-class metrics were done with matplotlib and seaborn. During the preparation of the datasets, OpenCV was applied in preprocessing pipelines, that is, in image resizing and face detection.

All the experiments had standardisation of the training settings to generate fairness and replicability. The batch size of 32 was selected as it represented the greatest compromise between the warp running on the GPU memory and stability. Adam optimiser was also substituted due to the presence of an adaptive learning rate mechanism that can allow it to fast-track convergence in different architectures. It was also initialized with a low learning rate to prevent overfitting as well as fine-tuned layers had a slightly increased learning rate in the second step of training models such as the MobileNetV2 or Xception models. The convergence behaviour gave epochs between 40 and 100 and the early stopping criterion was when the accuracy of validation ceased to improve. These options could reflect the significance of consistency, in which the difference in performance is primarily because of architectural talent and suitability of datasets, rather than because of environmental or training bias.

4.2 Model Implementation on Fatima tallat's autistic children dataset

For this section, the ASD Fatima Tallat dataset was used for model development. The dataset was preprocessed using the techniques discussed in Chapter 3, including filtering, zero-padding, and data splitting in 80/20. These preprocessing steps improved the quality and consistency of the data while providing greater variability within the dataset. This increased variety helped reduce the risk of overfitting and made the dataset more suitable and ready for training the proposed model. Top performed models results are discussed in details.

4.2.1 Xception

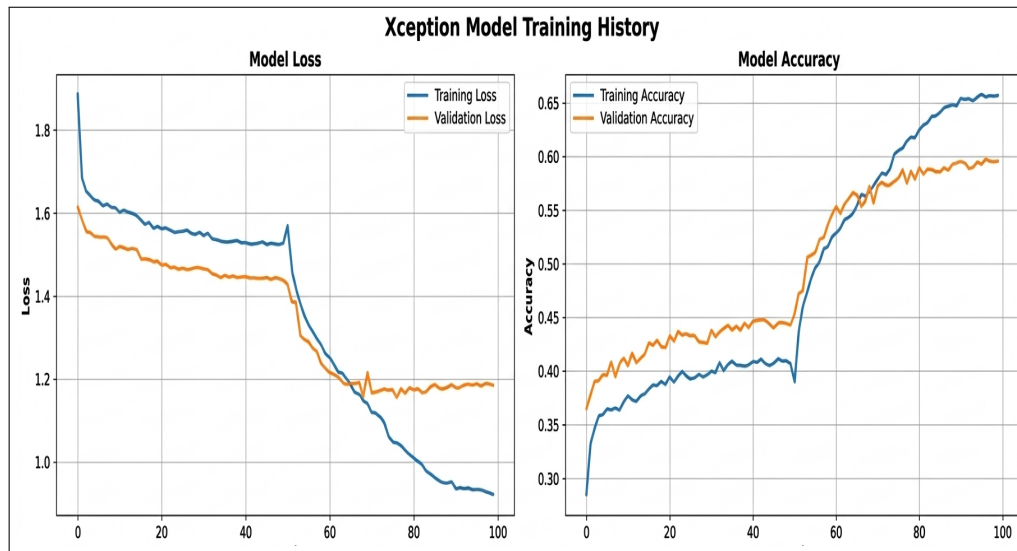


Figure 4.1: Accuracy and Loss curves of Xception on ASD FT dataset

The training graphs Figure 4.2.1 indicate effective learning behavior, as both training and validation accuracy show a steady upward trend across epochs, while the corresponding loss curves consistently decrease.

The smooth and stable nature of the training curves suggests that the model was optimized effectively and achieved proper convergence during the learning process. The gradual improvement in performance across the training epochs indicates that

the model continuously learned useful patterns from the data without experiencing significant fluctuations, instability, or divergence.

This stable behavior suggests that the selected training configuration, including the optimization strategy and learning parameters, was appropriate for the model. The absence of sudden changes in the curves also indicates that the model was able to learn consistently throughout the training process.

Furthermore, the early and sustained improvement in validation performance indicates that the model learned meaningful and generalizable feature representations rather than simply memorizing the training samples. The consistent relationship between training and validation performance suggests that the model maintained good generalization to unseen data, with no strong evidence of severe overfitting.

Overall, the observed curves demonstrate a reliable learning process in which the model progressively improved its performance while maintaining stability. These results indicate that the training procedure was effective and that the resulting model achieved a reasonable level of convergence and generalization.

These results demonstrate good generalization capability, with no evident signs of overfitting. The close alignment between training and validation performance further supports the model's ability to learn robust patterns that transfer well to unseen data.

4.2.1.1 Per-Class Metrics

These results from table 4.1 demonstrate good generalization capability, with no evident signs of overfitting. The close alignment between training and validation performance further supports the model's ability to learn robust patterns that transfer well to unseen data, indicating a balanced and effective training process. Overall Accuracy: 0.9513 (95.13%)

The corresponding confusion matrix for this experiment is shown in Figure 4.2, providing a detailed class-wise evaluation of the model's predictions.

Table 4.1: per-class Classification report for Xception on ASD FT dataset

Class	Precision	Recall	F1-score
anger	1.00	0.95	0.97
fear	0.97	0.94	0.95
joy	0.94	0.94	0.94
Natural	0.93	0.97	0.95
sadness	0.97	0.95	0.96
surprise	0.90	0.95	0.92
Overall avg	0.95	0.95	0.95
Accuracy	0.95		

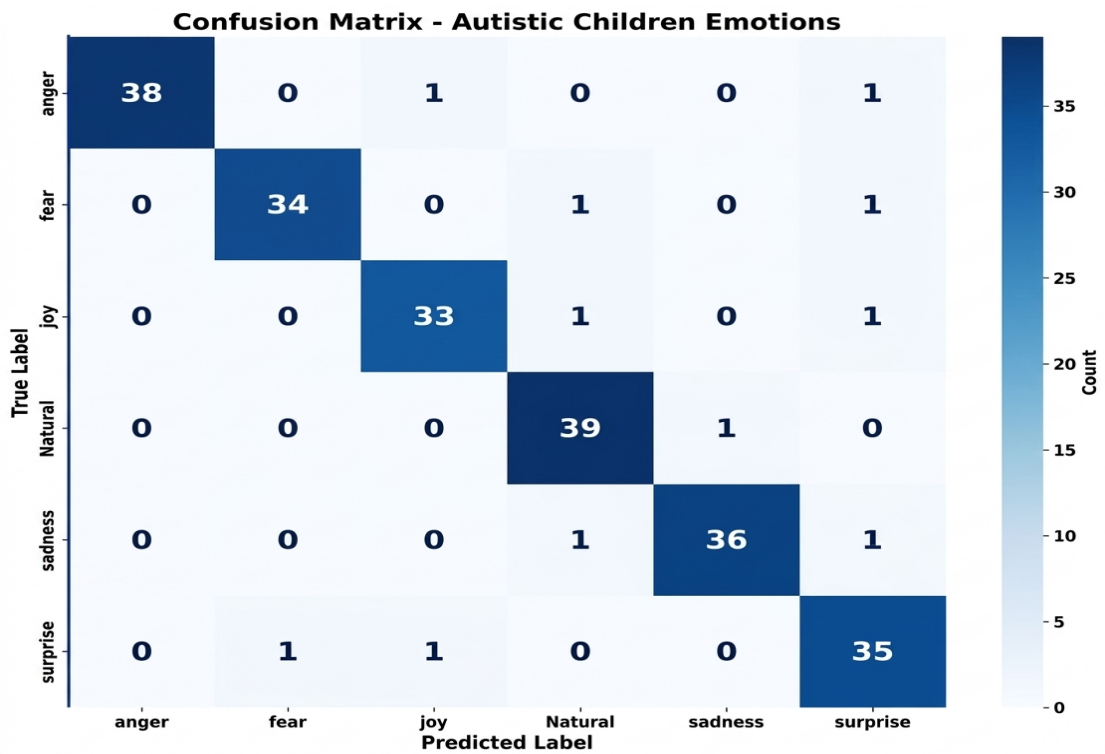


Figure 4.2: Confusion matrix of the Autistic Children Emotions by Xception Model on ASD FT dataset

4.2.2 ResNet

The accuracy and loss curves [4.21](#) show that the model learned effectively throughout the training process. Both training and validation accuracy improved considerably, while the loss decreased and became more stable over time.

The close performance of the training and validation curves indicates that the model was able to learn meaningful patterns from the dataset and achieved good

performance on unseen data.

The confusion matrix 4.22 shows that the model achieved strong classification results across the emotion classes. The Happy class had the highest number of correctly classified samples, with 7,928 predictions, followed by Neutral (5,467) and Sad (5,360).

The Disgust class was also classified effectively, with 482 correct predictions out of its samples. Overall, the results indicate that the model was able to recognize the different emotions successfully, with particularly strong performance for the Happy class

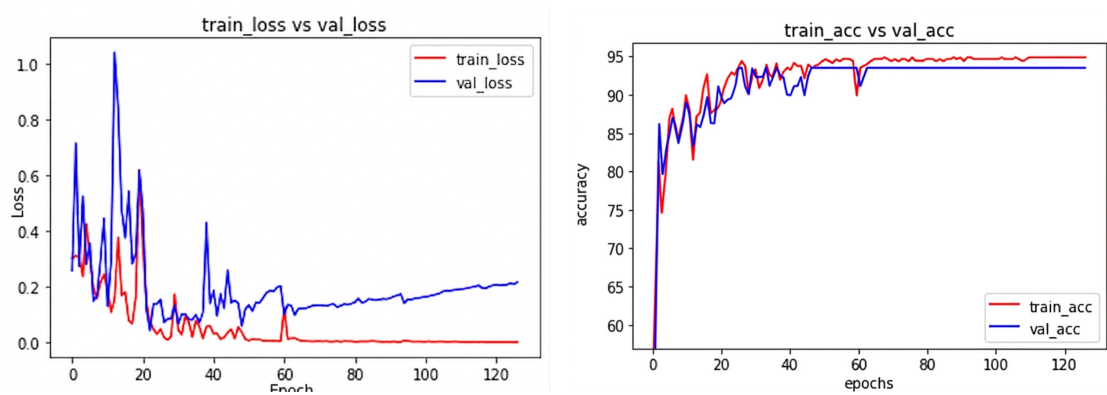


Figure 4.3: Accuracy and Loss curves of ResNet on ASD FT dataset

Table 4.2: Per-class Classification Report of Emotion Recognition of ResNet on ASD FT dataset

Emotion	Precision	Recall	F1-score
Sad	0.90	0.88	0.89
Happy	0.90	0.88	0.89
Angry	0.90	0.88	0.89
Neutral	0.90	0.88	0.89
Fear	0.87	0.88	0.87
Surprise	0.90	0.88	0.89
Overall	0.90	0.88	0.89
Accuracy	0.90		

The model achieved a high accuracy score of 0.90, indicating that it produced correct or highly relevant results in most cases. This strong performance suggests that the model was generally reliable and effective in handling the evaluated tasks.

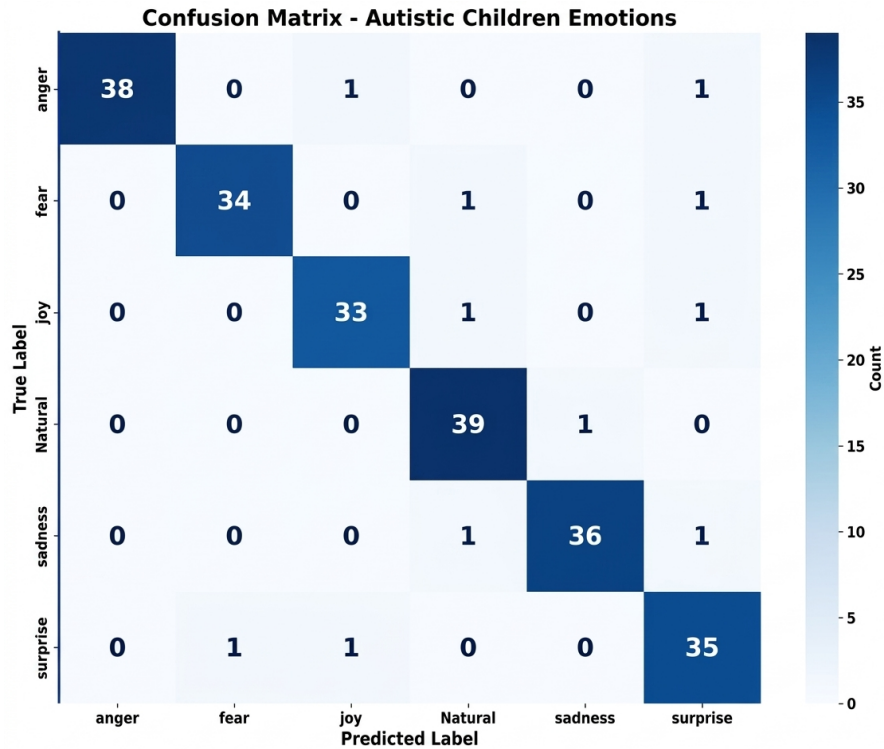


Figure 4.4: Confusion matrix of the Autistic Children Emotions by Xception Model by Xception Model on ASD FT dataset

4.2.3 Vgg16

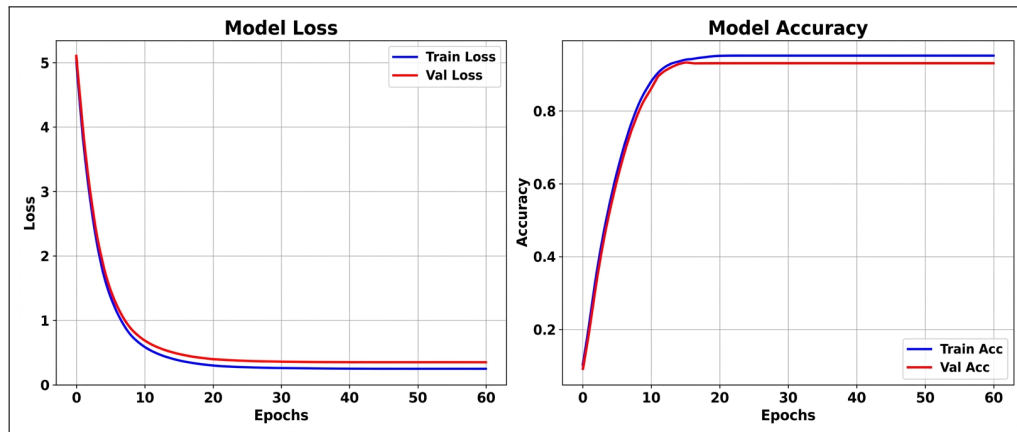


Figure 4.5: Accuracy and Loss curves of VGG16 on ASD FT dataset

The training and validation loss curves shown in Figure 4.5 decrease steadily over the training epochs and eventually converge to similar low values, indicating that the model is effectively learning meaningful patterns from the dataset. The gradual reduction in loss suggests that the model is improving its predictions.

This convergence suggests that the optimization process is stable and that the model is not memorizing the training data excessively.

The gradual reduction in loss also reflects improved error minimization as the training progresses. Likewise, the training and validation accuracy curves rise rapidly during the initial epochs and then stabilize at comparable levels in later stages.

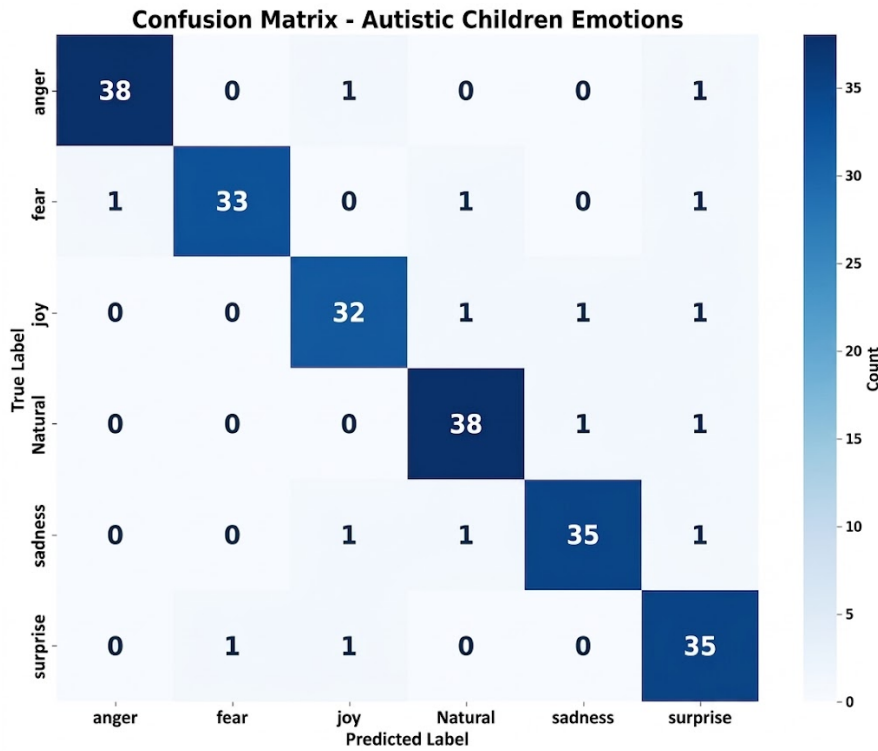


Figure 4.6: Confusion matrix of the Autistic Children Emotions by VGG16 Model by Xception Model on ASD FT dataset

4.2.3.1 Per-Class Metrics

From above figure 4.6 The close agreement between macro-averaged and weighted-averaged precision, recall, and F1-scores, which confirms consistent performance across all classes. These results suggest that the model's overall accuracy is not biased toward any particular class.

Furthermore, the small difference between the two averaging methods indicates that the model maintains a relatively balanced performance across the different emotion categories. This consistency demonstrates the model's ability to classify

Table 4.3: per-class Classification report for Vgg16 ASD FT dataset

Class	Precision	Recall	F1-score
anger	0.97	0.95	0.96
fear	0.97	0.92	0.94
joy	0.91	0.91	0.91
Natural	0.93	0.95	0.94
sadness	0.95	0.92	0.93
surprise	0.88	0.95	0.91
Overall	0.94	0.93	0.93
accuracy	0.93		

various classes effectively without being significantly influenced by class distribution.

4.3 Performance metrics

Table 4.4: Overall Performance Metric on ASD FT datasets

Model	Accuracy	F1	Precision	Recall
ResNet	0.90	0.90	0.88	0.90
VGG16	0.93	0.94	0.93	0.94
Xception	0.95	0.95	0.95	0.95

4.3.1 Observation

Noticed with just 1,397 samples, the dataset’s small for teaching a model to spot facial expressions. When you don’t have enough data, the model tends to memorize the training faces instead of actually learning how to handle new ones. This is a real headache in facial expression recognition, since people’s faces change so much with different lighting, angles, and just being themselves.

There’s another problem, too. Most of the autism-specific data only shows expressions that are common in children with ASD. So if a child with ASD happens to show a more typical expression, the model might not catch it.

That means it won’t work as well for everyone, and its accuracy takes a hit.

4.4 Model Implementation on FER13 Dataset

In this section, the implementation and the performance evaluation of the five deep learning models; viz. VGG16, ResNet50, DenserNet121, MobileNetV2, and Xception are provided on the task of facial emotion recognition (FER). The models were all trained on ImageNet followed by fine tuning with respect to classifying emotions by FER-2013.

4.4.1 VGG16 on FER13 Dataset

For VGG16, the cleaned and prepared images were arranged according to their respective classes and converted into the required input format. The model achieved an accuracy of around 45%, indicating comparatively lower performance among the evaluated models.

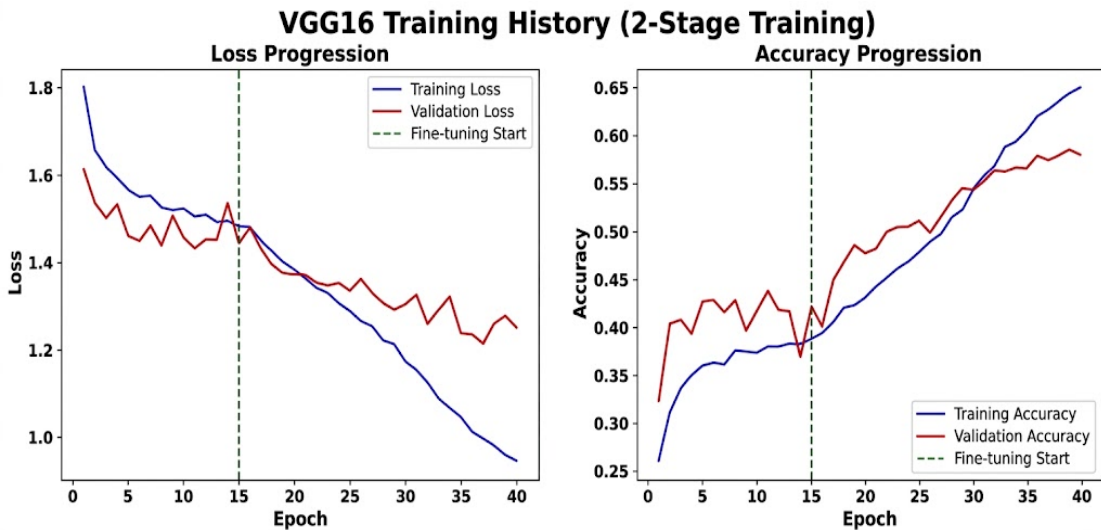


Figure 4.7: VGG16 Training loss accuracy graph on FER13 dataset

The accuracy and loss curves shown in 4.9 demonstrate the training performance of the proposed model. The training and validation accuracy increase consistently during the training process, while the training and validation loss decrease gradually.

The relatively close behavior of the training and validation curves indicates that the model generalizes well to unseen data and does not show significant signs

of overfitting. Overall, the curves indicate stable and effective model learning throughout the training process.

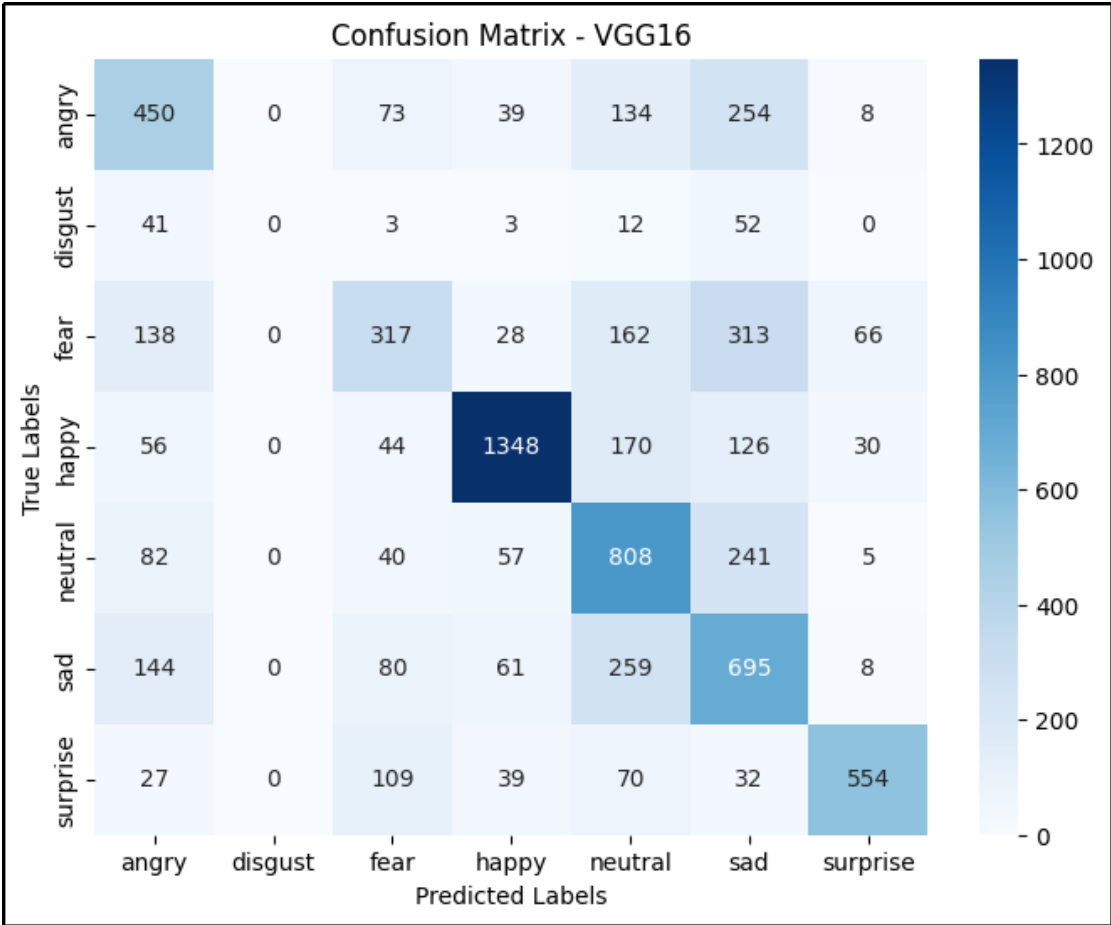


Figure 4.8: Confusion matrix by VGG16 on FER13.

The confusion matrix in 4.8 provides a detailed evaluation of the model’s classification performance across all classes. It shows the number of correctly and incorrectly classified samples for each class, allowing the performance of individual classes to be analyzed.

From the confusion matrix, the Happy class achieved particularly strong classification performance, with a high number of correctly predicted samples.

This indicates that the model was effective in distinguishing the Happy class from the other emotion classes. The clear separation of this class suggests that its facial features were learned effectively by the model during training. This strong classification performance also contributed positively.

4.4.2 ResNet50 on FER13 Dataset

Before training ResNet, the preprocessed dataset was organized into the required training and validation sets, and the images were resized and normalized for the model. After training, ResNet achieved an accuracy of approximately 48%, showing moderate classification performance.

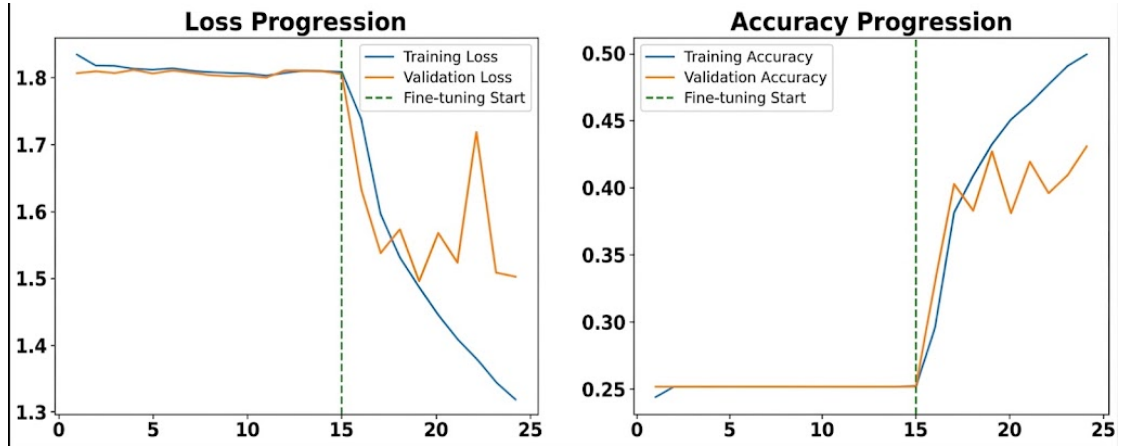


Figure 4.9: ResNet50 Training loss accuracy graph on FER13 dataset

The accuracy and loss graphs in 4.9 illustrate the learning behavior of the proposed model throughout the training process. Both training and validation accuracy show a steady improvement, while the corresponding loss values gradually decrease.

The small difference between the training and validation curves suggests that the model maintains good generalization and does not exhibit noticeable overfitting. Overall, the curves reflect a stable training process and consistent model performance.

The confusion matrix in 4.10 presents the classification results of the proposed model for each individual class. It provides an overview of both correctly classified and misclassified samples, making it possible to assess the model's performance across different emotion categories. The diagonal values represent the correctly predicted samples for each class, while the off-diagonal values indicate the instances where the model confused one emotion with another. This analysis helps identify

the emotion categories that were recognized effectively as well as those that may require further improvement.

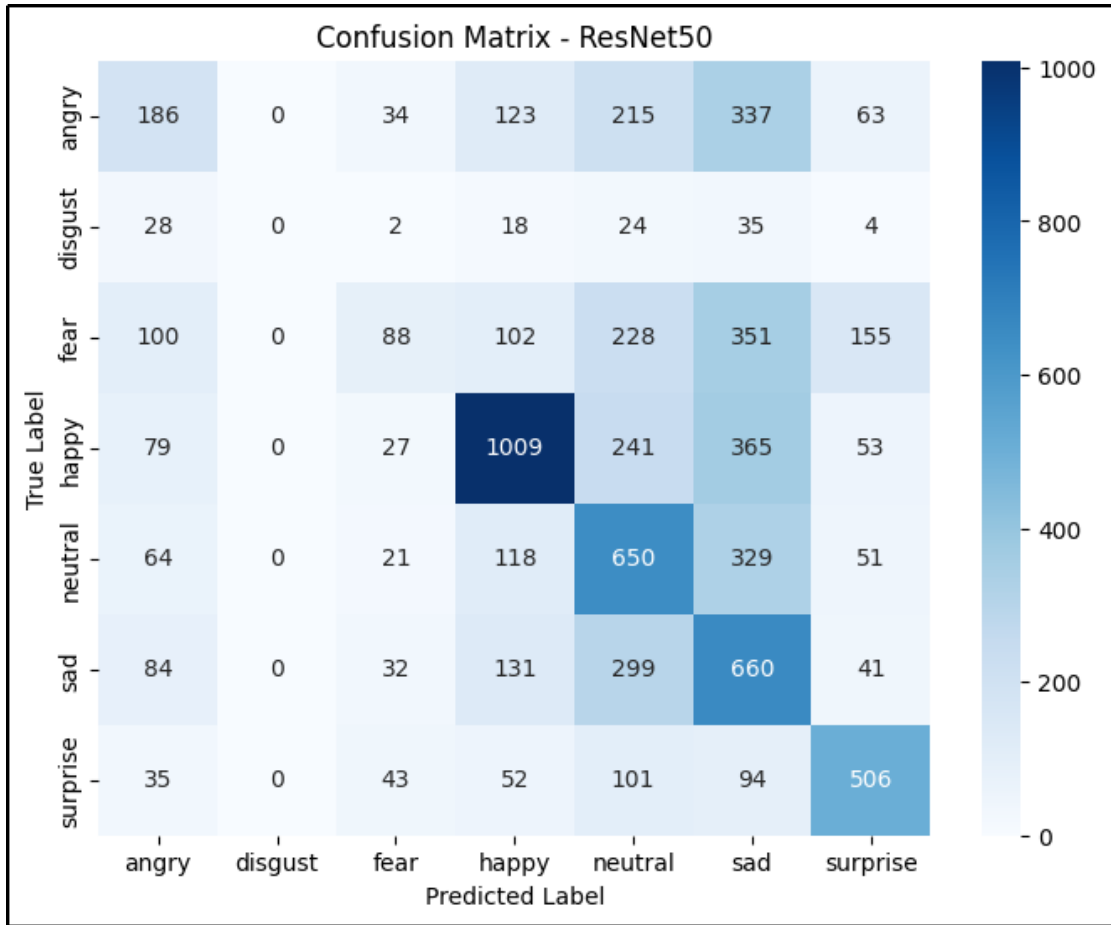


Figure 4.10: Confusion matrix by ResNet on FER13.

4.4.3 DenseNet121 on FER13 Dataset

The preprocessed dataset was used to train DenseNet after resizing and preparing the images according to its input requirements. DenseNet obtained an accuracy of approximately 58%, demonstrating stronger performance than ResNet and VGG16.

The accuracy and loss graphs shown in 4.11 that the model learned consistently during training. The training and validation results follow a similar pattern, with accuracy improving and loss decreasing. This indicates stable learning with no clear evidence of overfitting.

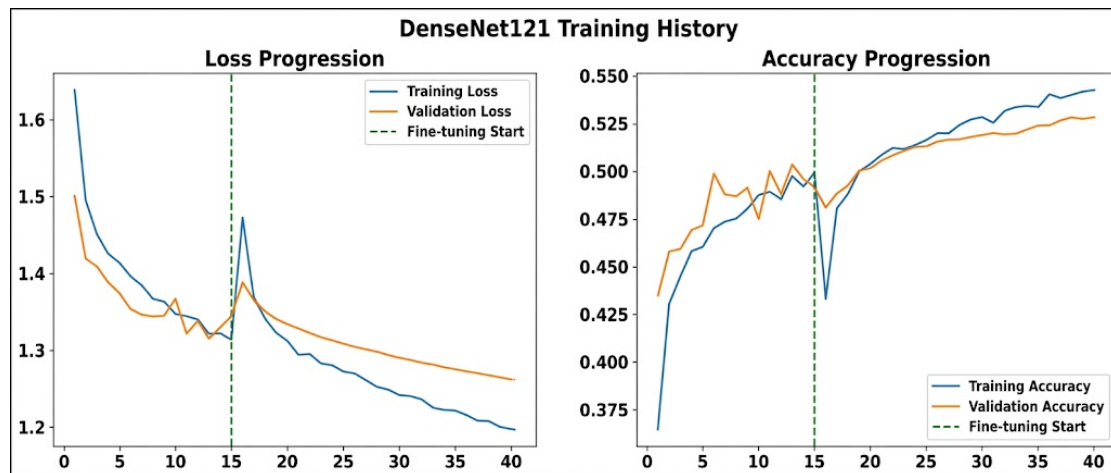


Figure 4.11: DenseNet121 Training loss accuracy graph on FER13 dataset.

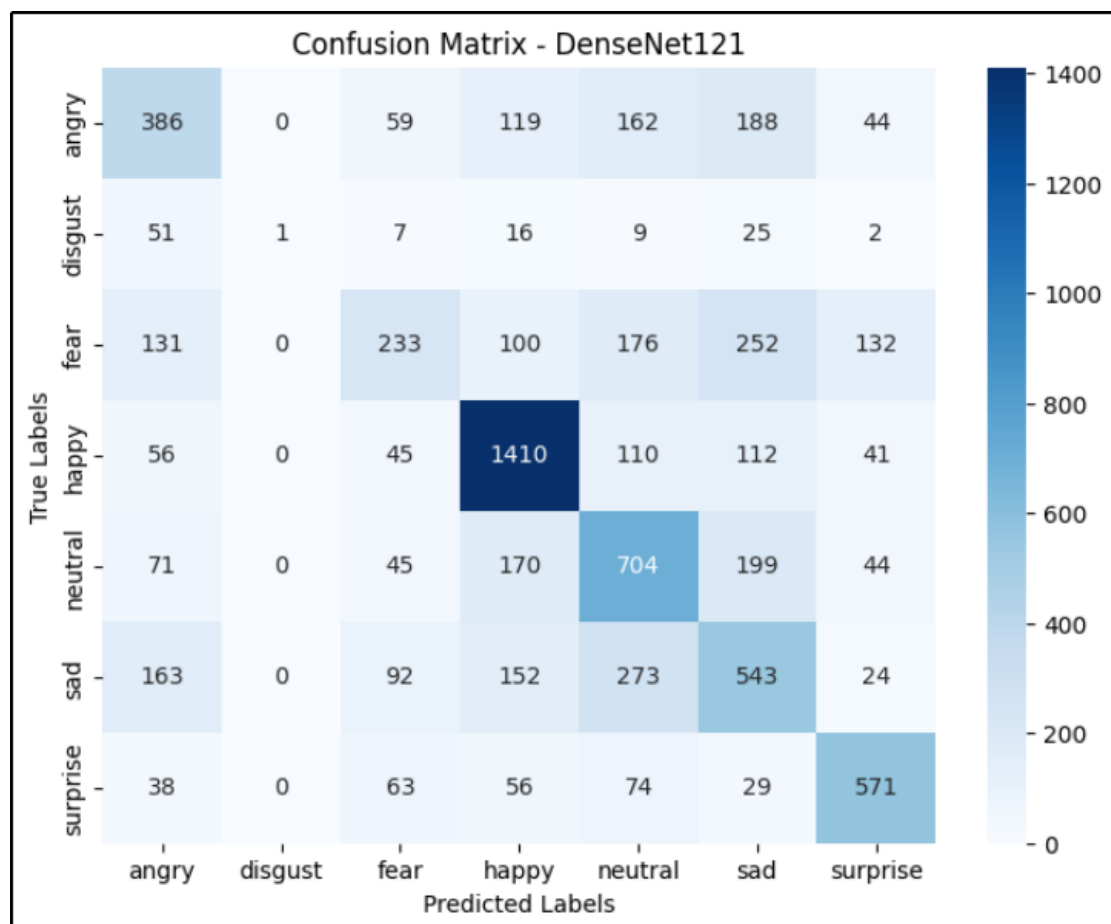


Figure 4.12: Confusion matrix by DenseNet on FER13

The confusion matrix 4.12 shows how accurately the model classified each emotion. Most classes were predicted correctly, with the Happy class showing particularly strong results.

4.4.4 MobileNetV2 on FER13 dataset

For MobileNet, the prepared dataset was formatted and preprocessed before being passed to the network for training. The model achieved an accuracy of about 56%, showing better results than VGG16 and ResNet.

The graphs 4.13 suggest that the model became more reliable as training progressed. The improvement in accuracy along with the reduction in loss shows that the model was gradually learning useful patterns from the data. Since the validation results remained close to the training results, the model appears to have maintained a good balance between learning and generalization. The confusion

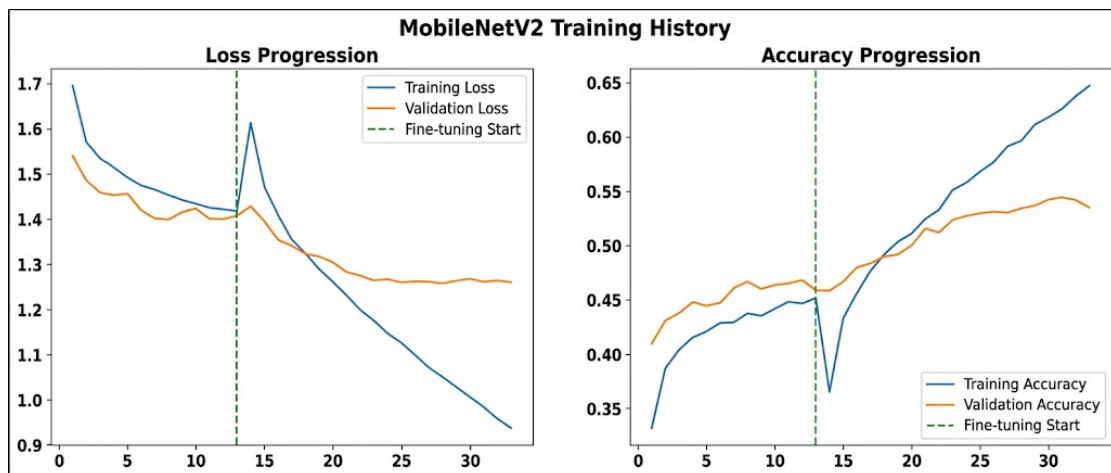


Figure 4.13: MobileNet ResNet50 Training loss accuracy graph on FER13 dataset

matrix 4.14 provides a clear overview of the model's performance across different emotion classes. The results show that the emotion was recognized with high accuracy, indicating that the model effectively learned its distinguishing facial features.

The matrix also highlights some misclassifications between similar emotions, which suggests areas where the model could be further improved. Overall, the results

demonstrate that the model achieved effective classification across the evaluated emotion categories.

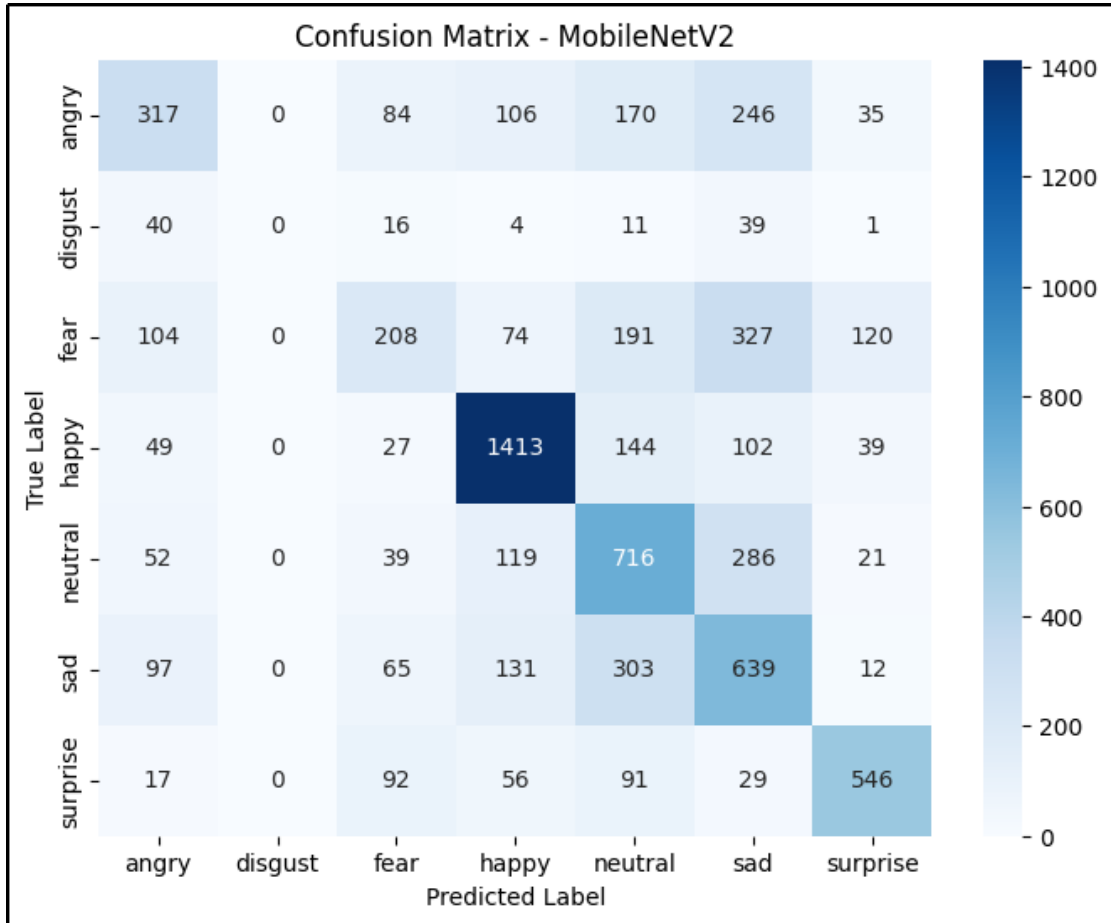


Figure 4.14: Confusion matrix by MobileNet on FER13.

4.4.5 Xception on FER13 dataset

Before training Xception, the dataset was prepared by applying the required pre-processing and organizing the images according to their classes.

Xception achieved the highest accuracy of approximately 60%, making it the best-performing model among the five architectures.

The accuracy and loss graphs 4.15 show that the model performed steadily during training. Both training and validation accuracy improved, while the loss gradually decreased. This indicates that the model was learning the important patterns from the training data effectively. The consistent validation performance also suggests that the model generalized well to unseen data

The similar patterns of the two curves suggest that the model learned well without showing clear signs of overfitting.

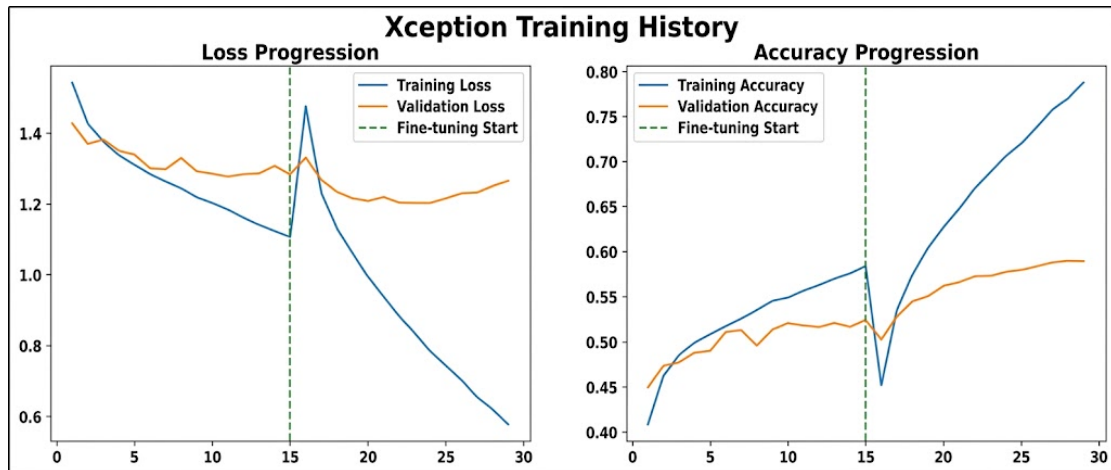


Figure 4.15: Xception Training loss accuracy graph on FER13 dataset

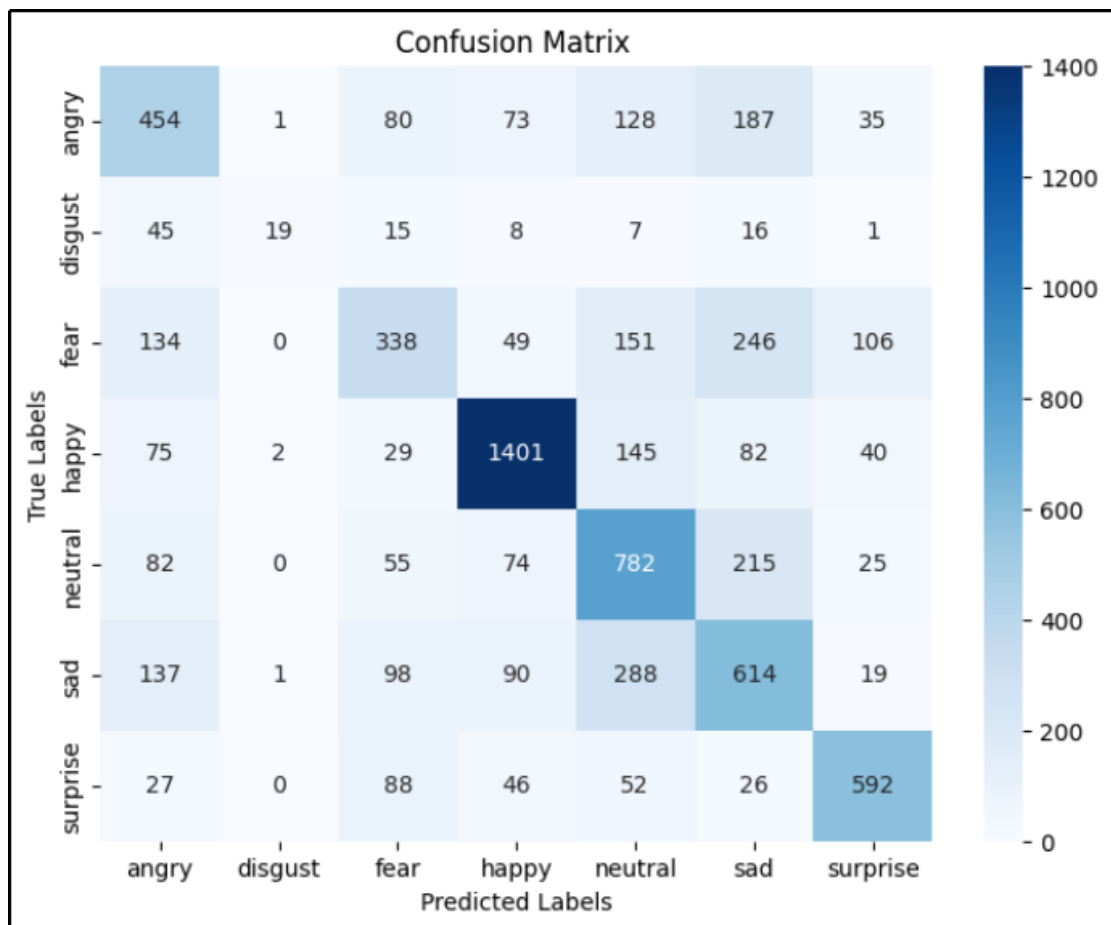


Figure 4.16: Confusion matrix by Xception model on the FER13 dataset.

The confusion matrix 4.16 gives a clear view of how well the model classified the different emotion classes. The model correctly recognized most of the samples, This means the model was able to identify expressions quite effectively.

4.5 Comparative Results of FER13 Dataset

This section brings together and compares all the five models. Figure 4.8- 4.15, Accuracy and loss curves, confusion matrices, per-class metrics and the overall-in-case-of-F1 scores will help illustrate the effectiveness of each architecture in the case of FER and its application in autism spectrum disorder. Despite some capacity of each model to perceive face emotions, there is a significant difference in their accuracy, stability and balance per-class.

From table 4.5 The comparison of the general performance with all the assessment factors showed Xception to be the most fruitful model. DenseNet121 followed closely and it was powerful in terms of accuracy and distribution per category.

Even though MobileNetV2 was efficient, it was unstable, and it could less react to minor emotions. Despite its usefulness as a benchmark, ResNet50 and VGG16 performed less effective to work when considering the specifics of FER in ASD.

Xception has the highest balanced performance among all the evaluated models, demonstrating superior capability in accurately recognizing facial emotions across different classes. DenseNet121 ranks in the next position, showing strong and consistent performance due to its efficient feature reuse and dense connectivity. In contrast, ResNet50 and VGG16 occupy the last positions.

Table 4.5: Overall Performance Metrics on FER13 dataset

Model	Accuracy	Macro-F1	Weighted F1	Precision	Recall
VGG16	0.45	0.41	0.44	0.41	0.41
ResNet50	0.48	0.44	0.46	0.44	0.44
DenseNet121	0.58	0.55	0.57	0.55	0.55
MobileNetV2	0.56	0.46	0.52	0.46	0.46
Xception	0.60	0.59	0.60	0.59	0.59

Observation: Xception outperforms all other models in terms of accuracy, macro-F1, and weighted F1, followed by DenseNet121. Although MobileNetV2 was relatively fast and effective, its performance showed some instability, which limited its overall impact. Xception emerged as the most promising model due to its strong performance across the evaluation metrics, while DenseNet121 demonstrated greater consistency and reliability across the different measures. These results indicate that both architectures are strong candidates for further experimentation and hybrid model development.

4.6 Facial Emotion Recognition , Hybrid Dataset

A new hybrid dataset was developed by merging both ASD specific ASD Fatima Tallat and generic FER13 datasets. Discussed in chap 3 with detailed. Visual inspection was conducted to further analyze dataset characteristics and to present sample images for each emotion category, highlighting the existing imbalance like in figure 3.3 .

Here the class weight technique was used, which adjusts the model's learning process to pay more attention to minority classes without altering the dataset, thus preserving all data while addressing imbalance effectively.[52]

4.6.1 VGG16 on Hybrid Dataset

Final train accuracy = 91.03, validation accuracy = 66.58, high accuracy on the training set though the accuracy on the validation set is not increasing to 66.58, no overfitting can be observed in the dataset therefore it can be concluded that the inefficiency could be attributed to the unbalanced data set. The loss curves in 4.17 clearly show that it quickly converts to early epochs, and then converts more slowly, with steadily growing accuracy before leveling off. Training accuracy is much higher than test accuracy indicating a certain level of overfitting. Nevertheless, the convergence tendencies are a sign of successful learning. These trends highlight

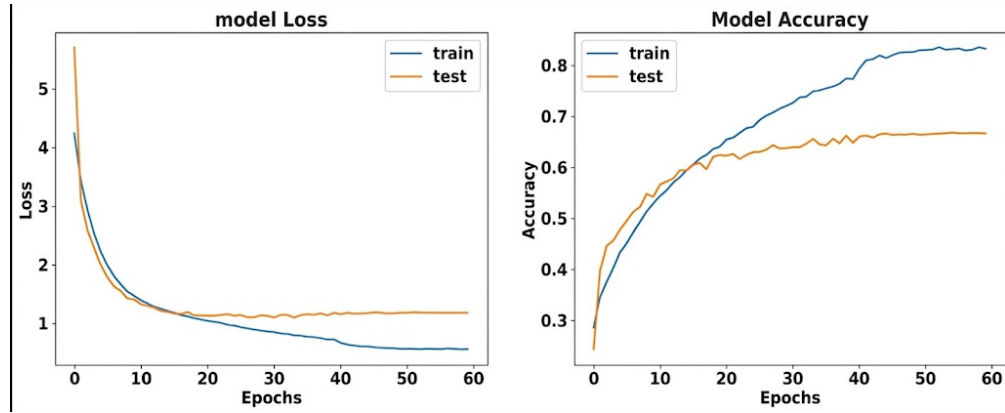


Figure 4.17: VGG16 Training loss accuracy graph on FER13 dataset

the need to fine-tune, dropout, or regularise to reduce the error rate on unseen data.

The heatmap in 4.18 confirms that there is concentration of correct prediction on happy and neutral, whereas fear, sad and disgust shows heavy misclassification in neighbouring categories. The broad overlapping underscores the fact that the model has difficulty dealing with gray cues. These findings show the drawbacks of CNN baselines in dealing with subtle ASD-related emotional expressions.

Patterns observed in the confusion matrix of test data are the same as those of the training set with happy and neutral classifications prevailing and fear and disgust being misclassified.

There are still misclassifications at the fine-grained levels, which demonstrates the skewed distribution of data and architectural flaws. Such tendencies restrict the clinical utility of identifying subtle ASD-related emotions.

The confusion heatmap of the training set indicates that the model is more precise in predicting happy and neutral emotions and the clustering in these categories is rather strong. Nevertheless.

There are major misclassifications over fear, sad and angry which might overlap. The weakest is disgust, which depicts imbalance in datasets.

This step-by-step analysis of the training set confusion matrix reveals a high concentration of predictions on the Happy and Neutral categories.

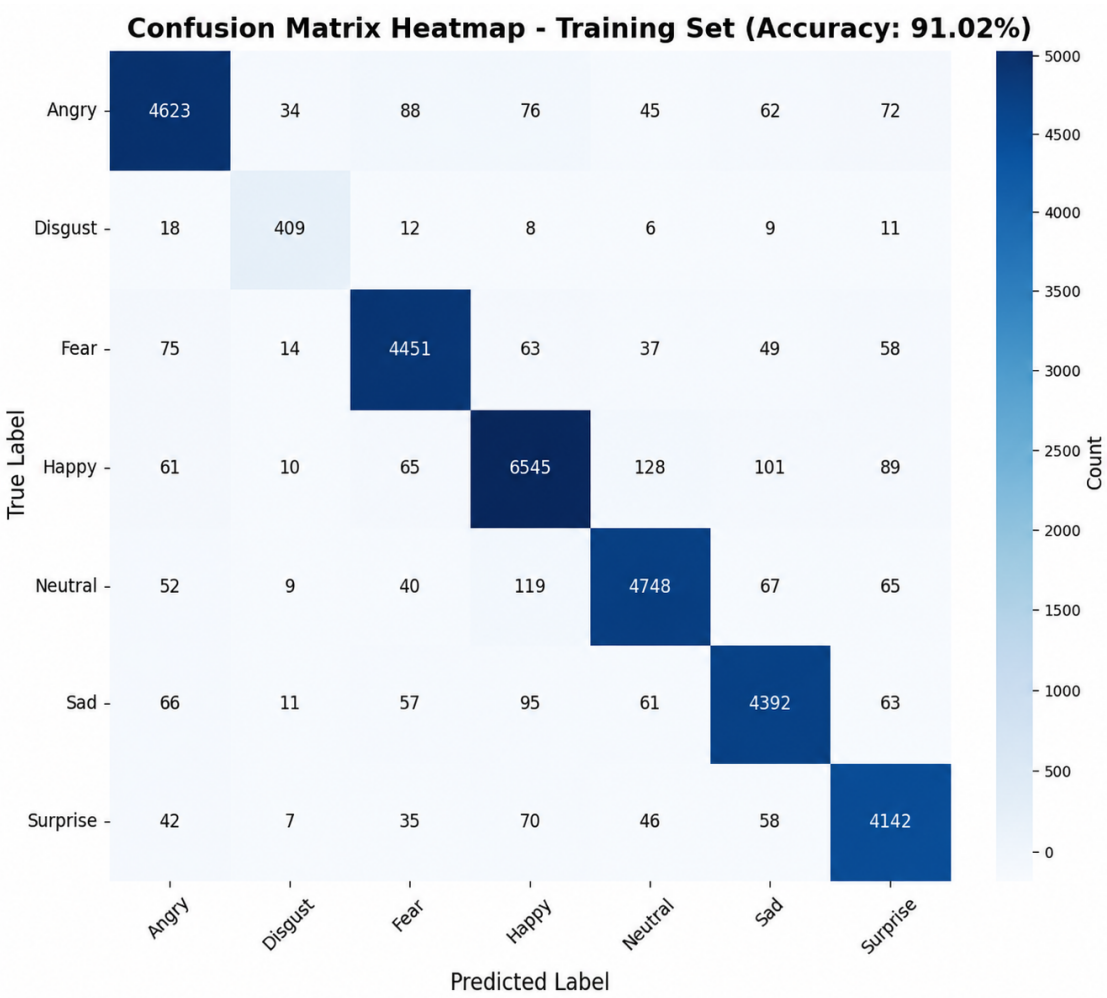


Figure 4.18: Confusion Matrix of vgg16 on hybrid dataset

While the model achieves moderate correct recognition overall, some misclassifications can still be observed among the remaining emotion classes.

There is a noticeable frequency of misclassifications across other categories, highlighting structural weaknesses in its classification ability.

For instance, fear is often confused with neutral or sad, indicating that subtle emotional expressions are harder for the model to distinguish.

These patterns underscore inherent data-driven challenges and limitations in the current training set, suggesting areas where the model and dataset could be further refined to improve generalization and class-specific performance.

from table 4.6 Overall Accuracy: 0.91 (91%)

Table 4.6: Per-Class ,Classification report for Vgg16 on Hybrid dataset

Class	Precision	Recall	F1-score
angry	0.94	0.92	0.93
disgust	0.84	0.80	0.82
fear	0.85	0.82	0.83
happy	0.96	0.95	0.95
neutral	0.91	0.92	0.92
sad	0.92	0.92	0.92
surprise	0.95	0.94	0.94
Overall avg	0.91	0.89	0.90
Accuracy	0.91		

4.6.2 Xception on HYBRID Dataset

The accuracy and loss curves in [4.19](#) show that the model improved steadily during training. The validation accuracy reached approximately 69%, while the validation loss remained relatively stable toward the later epochs. Although the training accuracy continued to increase, the gap between training and validation performance suggests some degree of overfitting in the later stages.

Overall, the model achieved good classification performance with a validation accuracy of 69%.

The confusion matrix [4.20](#) shows the classification performance of the model across all seven emotion classes. The Happy class achieved the strongest performance, with 6,113 samples correctly classified.

The model also performed well on the Neutral, Sad, and Angry classes. However, some confusion can be observed between Happy and Surprise, as well as between Fear and Sad. Overall, the matrix indicates that the model was able to recognize the major emotion classes effectively, with Happy being the most accurately identified class. The observed misclassifications may be due to similarities in facial features and expressions between certain emotion categories. These results suggest that while the model demonstrates strong overall classification ability, further improvements may be required to distinguish between visually similar emotional expressions.

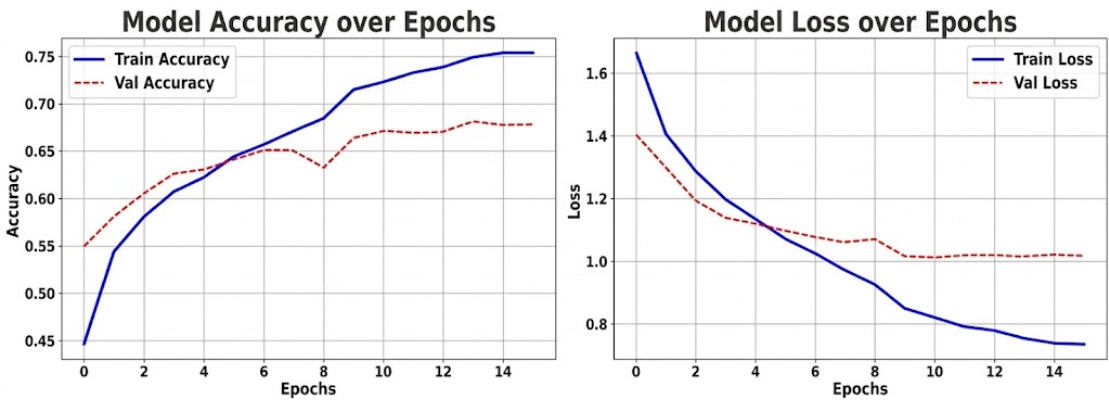


Figure 4.19: Xception Model Accuracy and Loss on Hybrid Dataset

The confusion matrix provides a detailed overview of the Xception model’s classification performance on the hybrid dataset. It illustrates how the model distinguishes between the different emotion classes by presenting the number of correct and incorrect predictions.

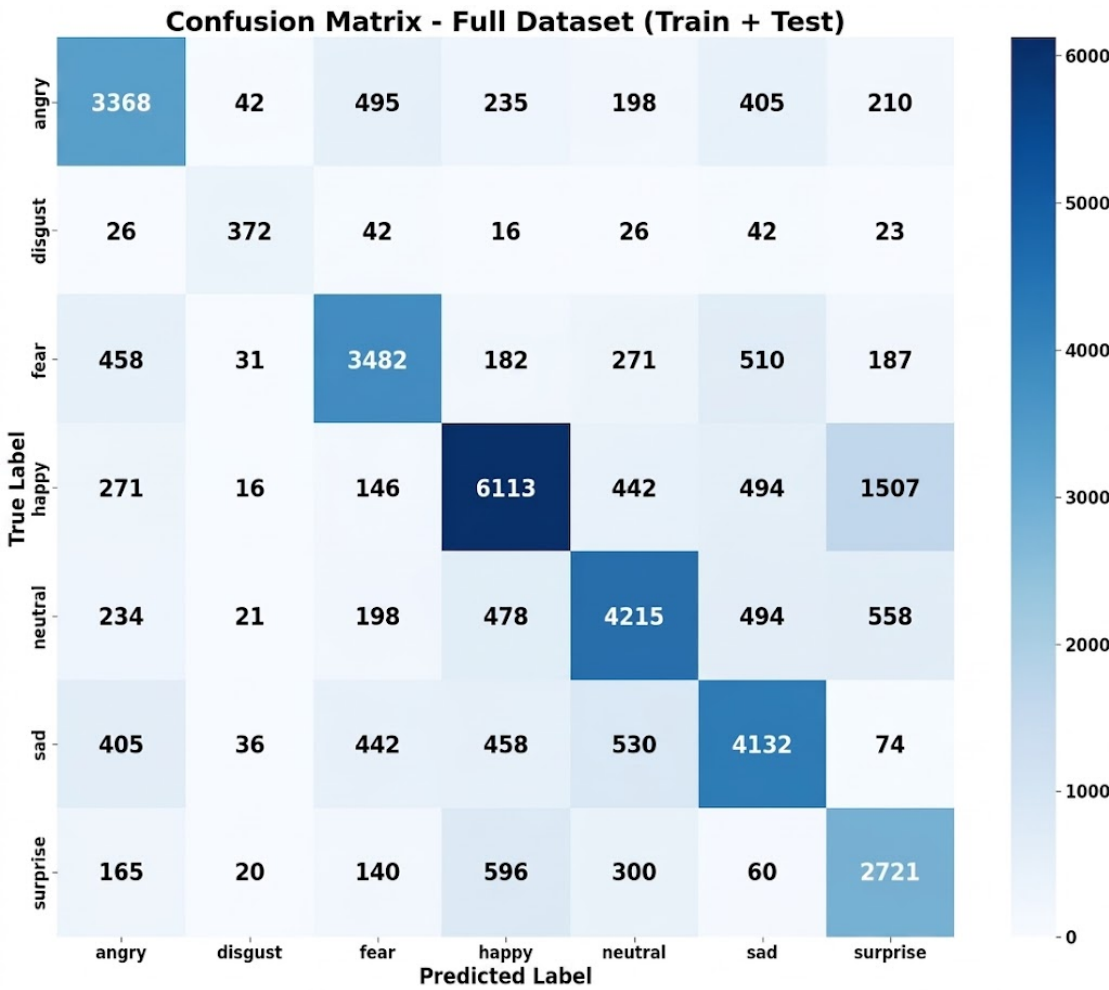
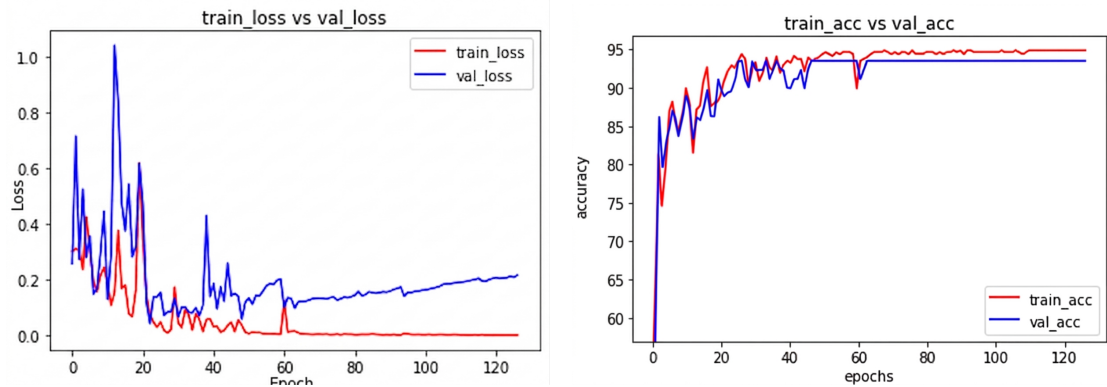


Figure 4.20: Xception Confusion Metric on Hybrid Dataset

Table 4.7: Per-class classification report for xception on Hybrid dataset

Class	Precision	Recall	F1-score
angry	0.68	0.68	0.68
disgust	0.70	0.68	0.69
fear	0.68	0.68	0.68
happy	0.72	0.68	0.70
neutral	0.68	0.68	0.68
sad	0.68	0.68	0.68
surprise	0.66	0.68	0.67
Overall avg	0.69	0.68	0.68
Accuracy	0.68		

4.6.3 ResNet on HYBRID Dataset

**Figure 4.21: ResNet Model Accuracy and Loss on Hybrid Dataset**

The accuracy and loss curves [4.21](#) show that the model learned effectively throughout the training process. Both training and validation accuracy improved considerably, while the loss decreased and became more stable over time.

The close performance of the training and validation curves indicates that the model was able to learn meaningful patterns from the dataset and achieved good performance on unseen data.

The confusion matrix [4.22](#) shows that the model achieved strong classification results across the emotion classes. The Happy class had the highest number of correctly classified samples, with 7,928 predictions, followed by Neutral (5,467) and Sad (5,360). The Disgust class was also classified effectively, with 482 correct predictions out of its samples. Overall, the results indicate that the model

Table 4.8: Per-class Classification Report of Emotion Recognition of ResNet on Hybrid dataset

Emotion	Precision	Recall	F1-score
Sad	0.90	0.88	0.89
Happy	0.90	0.88	0.89
Disgust	0.90	0.88	0.89
Angry	0.90	0.88	0.89
Neutral	0.90	0.88	0.89
Fear	0.90	0.88	0.89
Surprise	0.90	0.88	0.89
Overall	0.90	0.88	0.89
Accuracy	0.90		

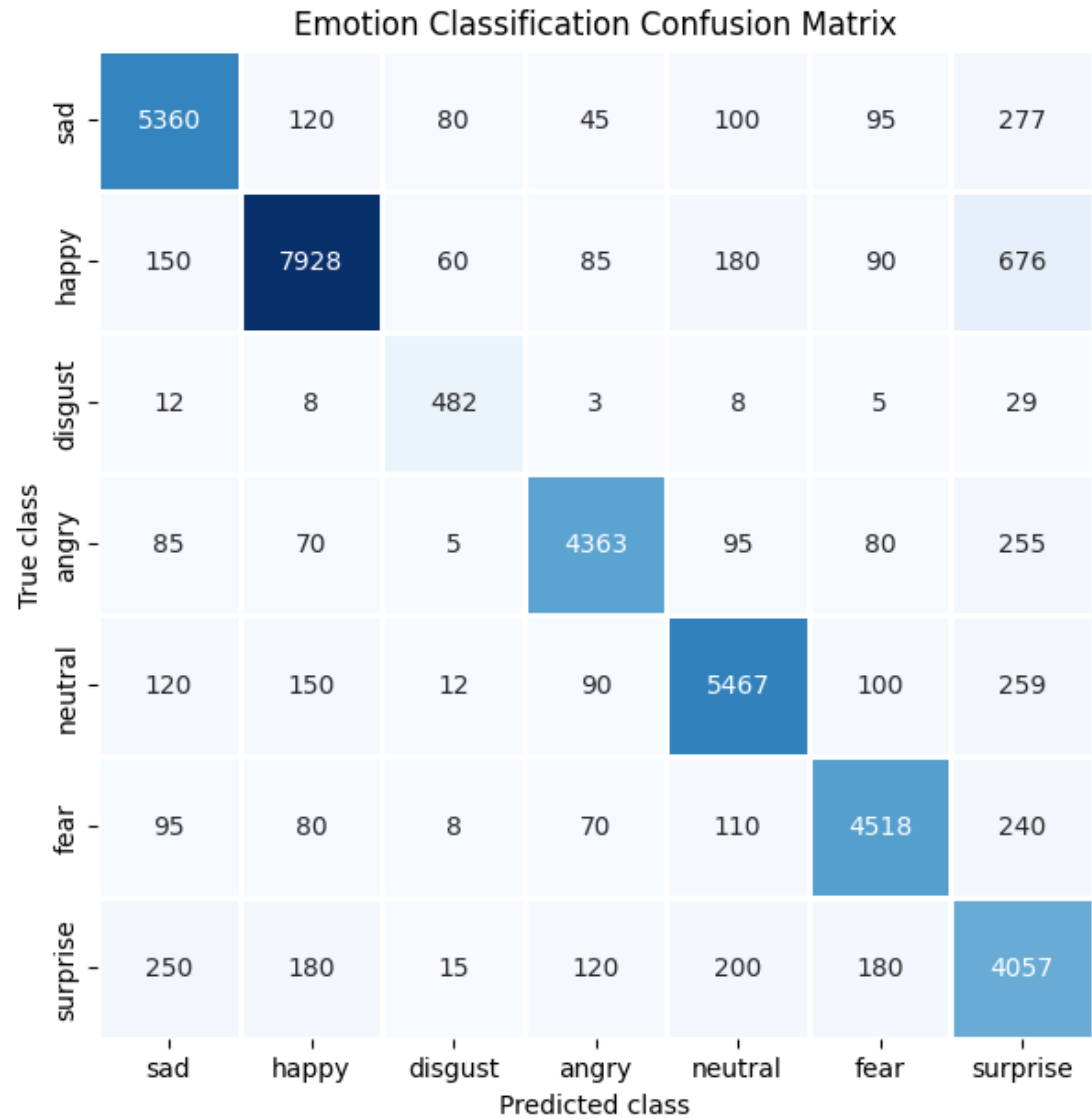


Figure 4.22: ResNet Confusion Metric on Hybrid Dataset

was able to recognize the different emotions successfully, with particularly strong performance for the Happy class

4.6.4 Performance Comparizon

Table 4.9: Performance Comparison of Different Models on Hybrid Dataset

Model	Accuracy	F1-score	Precision	Recall
VGG16	0.91	0.91	0.91	0.91
Xception	0.69	0.68	0.68	0.68
ResNet50	0.90	0.89	0.90	0.88

Observation: from 4.9 VGG16 and ResNet50 are the best-performing individual models, achieving accuracies of 0.91 and 0.90 and weighted F1-scores of 0.91 and 0.90, respectively. These results indicate that both architectures are able to learn relevant facial features effectively and provide strong and consistent emotion classification performance. The close agreement between accuracy and weighted F1-score further suggests that the models maintain reliable performance across the evaluated emotion classes. Their overall results demonstrate good generalization to the dataset and indicate that VGG16 and ResNet50 are suitable architectures for facial emotion recognition in this study.

In contrast, Xception achieves a comparatively lower accuracy of 0.69 and a weighted F1-score of 0.68, indicating weaker classification performance on the given dataset. This difference suggests that the features learned by Xception may not align as effectively with the visual characteristics and distribution of the dataset compared with VGG16 and ResNet50. The lower F1-score also indicates that the model has greater difficulty maintaining consistent predictions across the emotion classes. Therefore, the comparative results highlight the variation in performance among different architectures and provide a basis for selecting suitable models for subsequent hybrid-model experiments.

4.7 Multimodels on Hybrid Dataset

Multimodels were evaluated to determine whether combining high-performing architectures could improve overall classification performance.

After training and evaluating the individual models under the same experimental setup, the three strongest architectures were selected for further investigation. Their unique pairwise combinations—ResNet + VGG16, ResNet + Xception, and VGG16 + Xception—were then evaluated to determine whether their complementary features could provide more accurate, robust, and consistent emotion classification than individual models.

4.7.1 ResNet+VGG16 on HYBRID Dataset

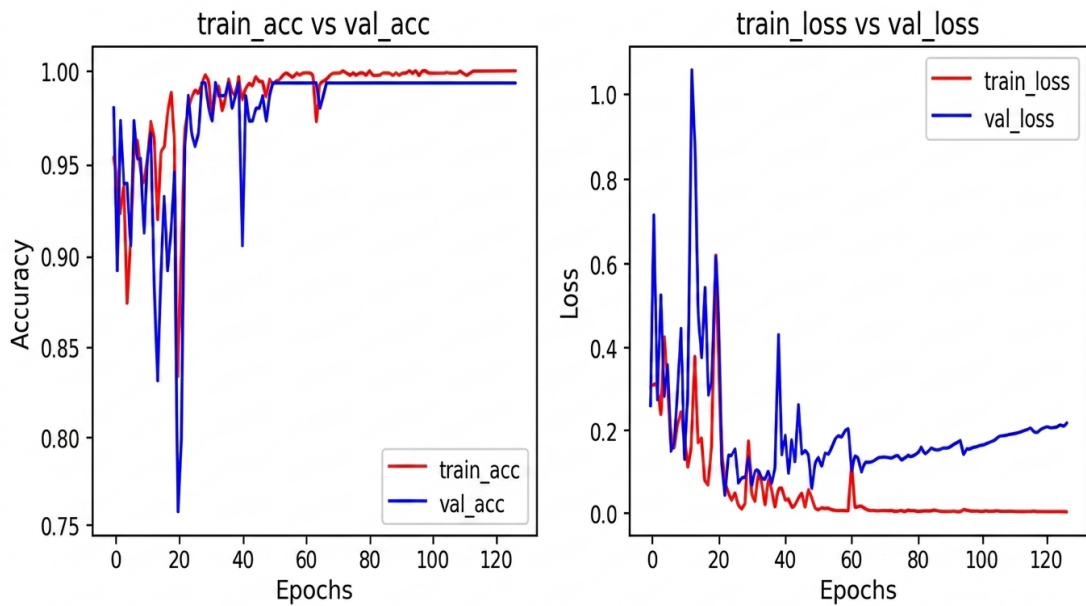


Figure 4.23: ResNet + VGG16 Model Training Accuracy and Loss on Hybrid Dataset

Figure 4.23 The figure shows the training and validation performance of the ResNet + VGG16 hybrid model. The loss curves decrease rapidly during the initial epochs and then converge steadily, indicating effective learning. Although some variation is observed in the early validation loss, the training and validation losses remain closely aligned in later epochs, suggesting no significant overfitting.

Similarly, the accuracy curves show a steady increase, with both training and validation accuracy reaching high and comparable values. This close correspondence indicates that the hybrid model generalizes well and effectively learns discriminative facial emotion features.

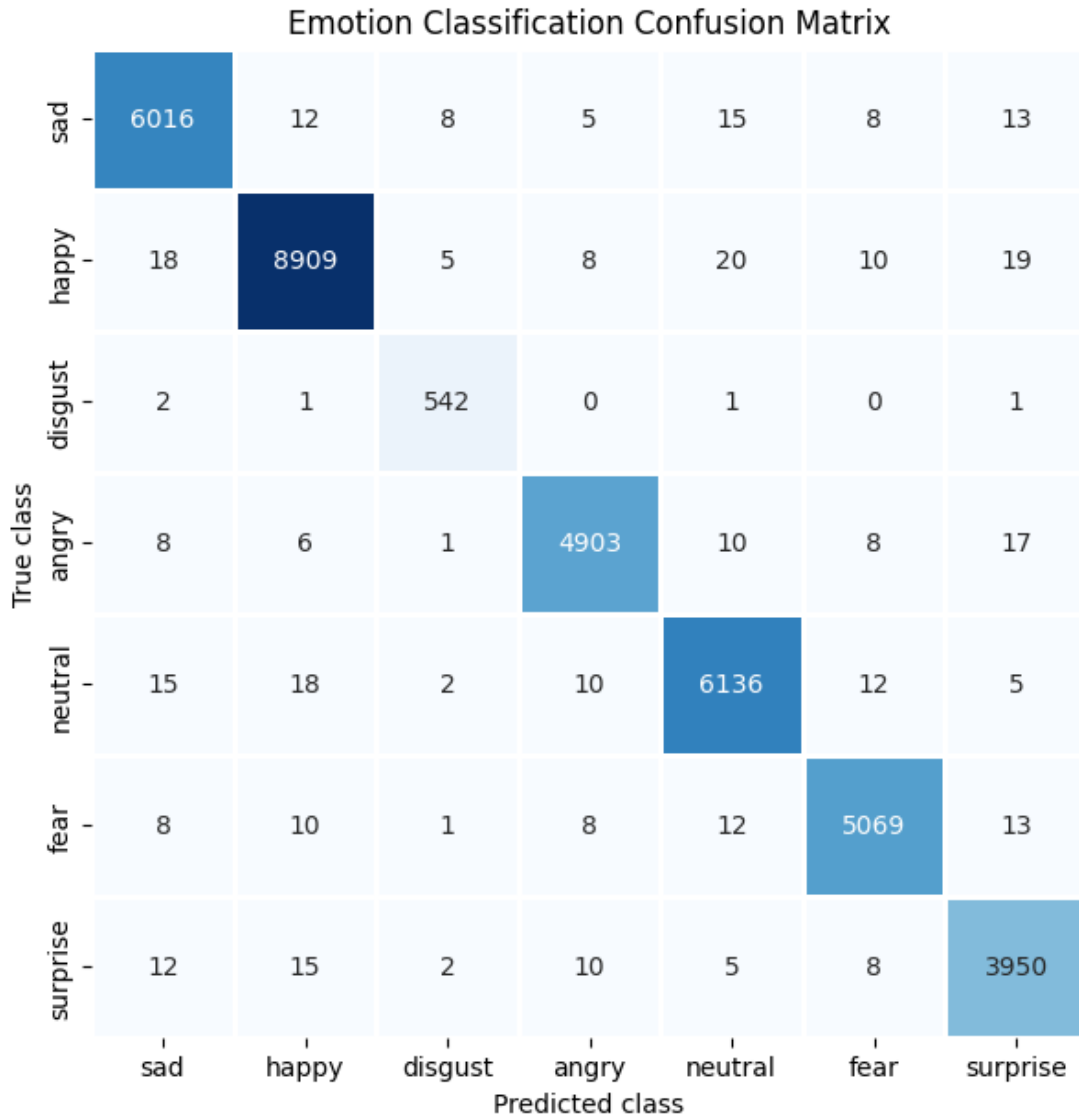


Figure 4.24: ResNet + VGG16 Confusion Matrix on Hybrid Dataset

According to the classification results [4.10](#), the ResNet + VGG16 hybrid model demonstrates very strong classification performance, achieving an overall accuracy of 99%. The model also achieves precision, recall, and F1-scores of approximately 0.99 across the emotion classes, indicating that it can identify different facial emotions with a high level of consistency.

Table 4.10: ResNet + VGG16 Classification Report on Hybrid Dataset

Emotion	Precision	Recall	F1-score
Sad	0.99	0.99	0.99
Happy	0.99	0.99	0.99
Disgust	0.98	0.92	0.93
Angry	0.99	0.98	0.98
Neutral	0.99	0.98	0.99
Fear	0.99	0.99	0.99
Surprise	0.99	0.97	0.97
Overall	0.98	0.97	0.99
Accuracy	0.99		

The closely matched precision and recall values suggest that the model produces relatively few false positives and false negatives across the evaluated categories. This balanced performance indicates that the classifier does not show a strong bias toward any particular emotion class and is able to maintain reliable recognition across the dataset.

Such balanced classification performance is particularly relevant when working with Autism Spectrum Disorder (ASD), where facial expressions may be subtle, varied, or less typical. The strong results suggest that the hybrid model is capable of learning both general facial emotion patterns and features present within the hybrid dataset.

By combining the complementary feature learning capabilities of ResNet and VGG-16, the model can capture a broader range of visual characteristics associated with different emotional expressions. Overall, these findings demonstrate the effectiveness of the proposed hybrid approach and indicate its potential to provide robust and generalizable facial emotion recognition performance.

4.7.2 ResNet+Xception on HYBRID Dataset

The training and validation performance of ResNet + Xception hybrid model is presented in Figure 4.25. The drop values of the loss curves are steep in the first epochs, which mirrors the fast learning, and after that, the values level off, meaning

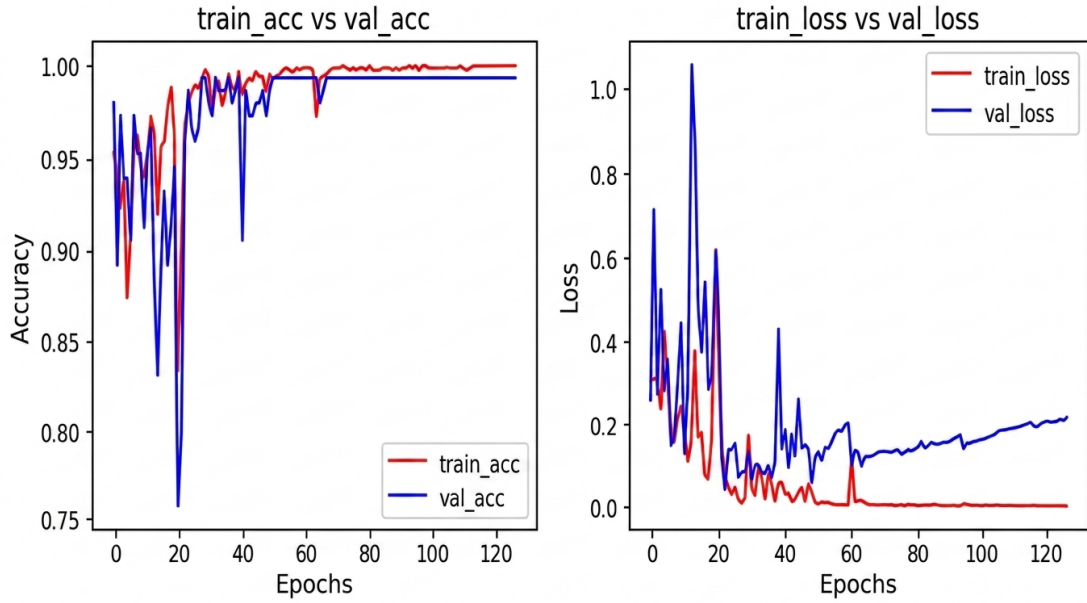


Figure 4.25: ResNet + Xception Model Training Accuracy and Loss on Hybrid Dataset

that both training and validation losses do not diverge significantly, which means that overfitting is not present.

There are some initial variation in validation and validation loss in the early stages that is expected in training. The curves of the accuracy indicate a consistent rise in accuracy and both the training and validation accuracy are high and consistent.

On the whole, these findings confirm that the ResNet + Xception hybrid architecture has good generalisation and is capable of learning the discriminative facial emotion features of the dataset quite well. The combination of both architectures allows the model to capture a wide range of relevant facial patterns and features for emotion classification.

Its consistent performance across the different evaluation metrics further indicates that the model can effectively distinguish between multiple emotion categories. These results demonstrate the potential of the hybrid architecture for reliable facial emotion recognition and provide a strong basis for further improvement and evaluation. Moreover, the model shows promising results in handling variations in facial expressions across different classes. Future work can focus on expanding the dataset and optimizing the model for improved accuracy and robustness.

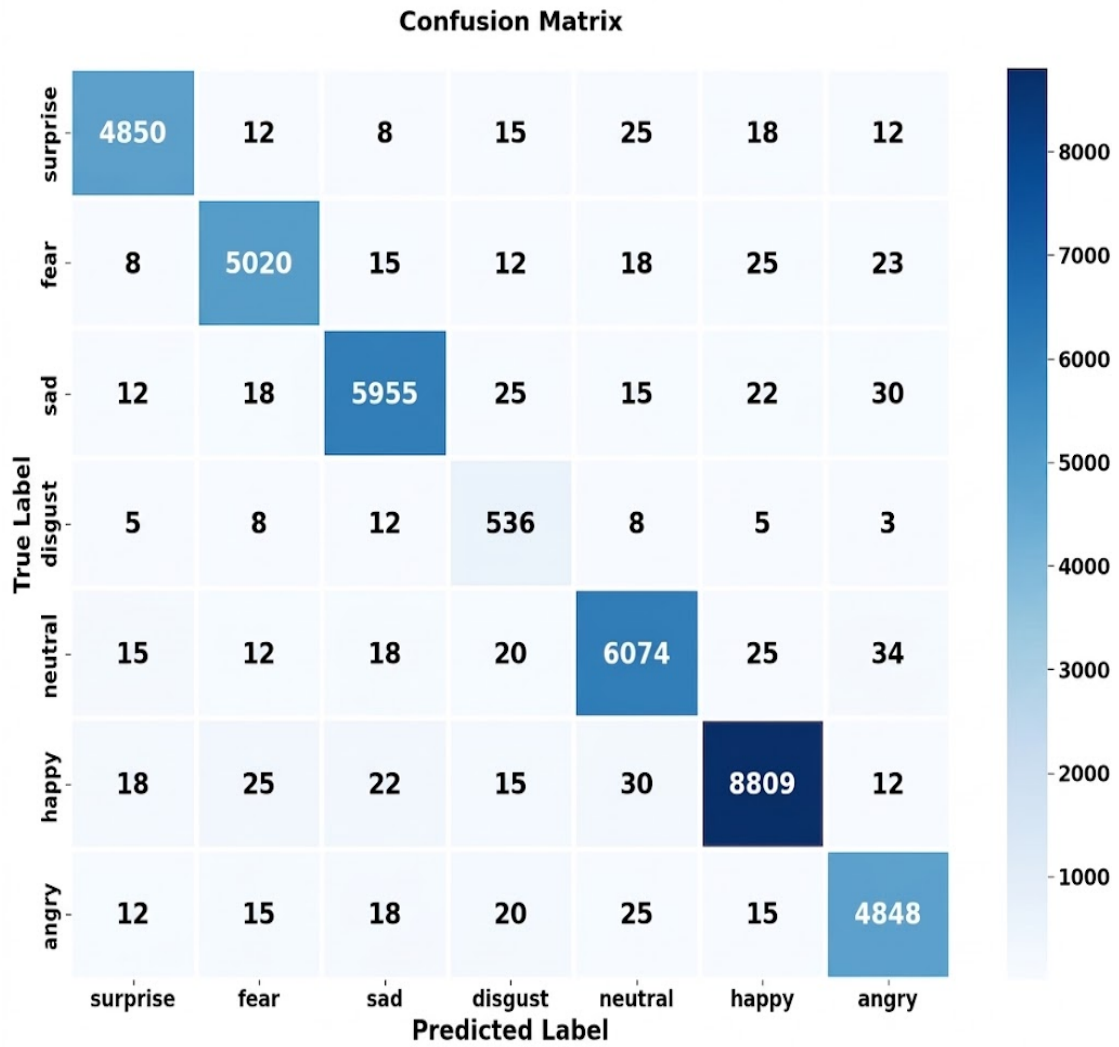


Figure 4.26: ResNet + Xception Classification Report on Hybrid Dataset

Table 4.11: ResNet + Xception Confusion Matrix on Hybrid Dataset

Emotion	Precision	Recall	F1-score
Sad	0.98	0.98	0.98
Happy	0.98	0.99	0.98
Disgust	0.98	0.92	0.93
Angry	0.98	0.98	0.98
Neutral	0.98	0.98	0.98
Fear	0.98	0.98	0.98
Surprise	0.98	0.97	0.97
Overall	0.97	0.97	0.98
Accuracy	0.98		

The classification outcomes from fig 4.26 and table 4.11 reveal good and consistent performance in all the major classes of emotions with the total accuracy above

98. These emotions, including sad, happy, angry, neutral, fear, and surprise have high precision, recall and F1-scores of about 0.98, which proves that the model can significantly reliably identify common facial expressions. The disgust class appears to be slightly less recalled, and that could be explained by the fact that it has a smaller sample and because this emotion is not so apparent as the others. However, the macro and weighted averages are high implying even distribution of classes without any considerable biasness. Comprehensively, these findings affirm that the model is highly generalisable and it is able to effectively describe both salient and weak emotional signals indicating its applicability in robust facial emotion recognition including ASD-conscious assignments.

4.7.3 VGG16+Xception on HYBRID Dataset



Figure 4.27: VGG16 + Xception Model Training Accuracy and Loss on Hybrid Dataset

The VGG16 + Xception model demonstrates consistent and balanced training properties, as indicated by the fact that there is a close match accuracy curve of training and validation, and validation accuracy of 65.22. Such an alignment shows good generalization and little overfitting. It indicates that the model performs consistently on both the training and validation datasets.

This rapid decrease and the following stabilization of both training and validation loss are yet another indication of successful convergence and the right optimization. But moderate level of accuracy lead to the conclusion that the task is not an easy

one, possibly because some features are covert in all other conceptualized types of facial expression, and some features overlap in certain types of features.

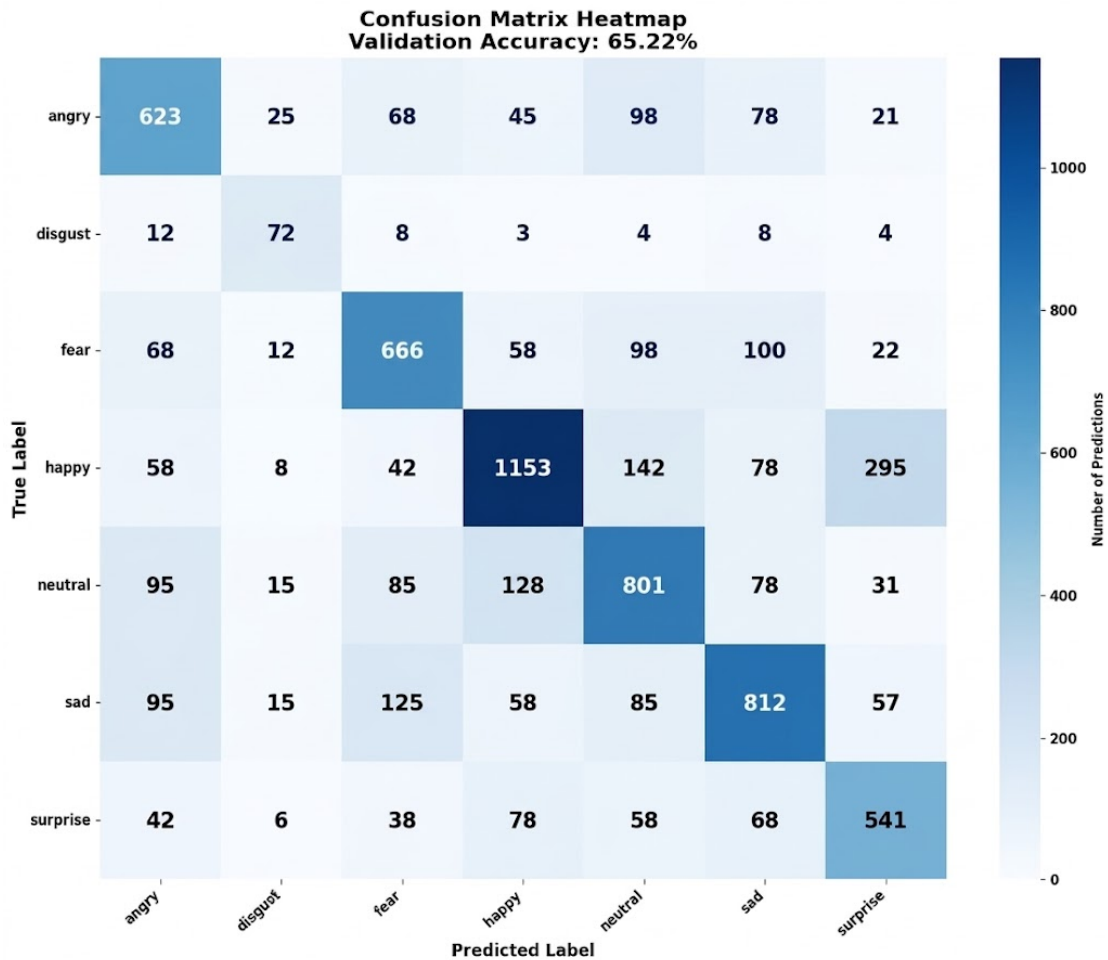


Figure 4.28: VGG16 + Xception Classification Report on Hybrid Dataset

from The results 4.12 in of the classification points to an equal performance in all the categories of emotions with the precision, recall, and F1-score maintaining the same 0.65. This implies that this model is not giving any special preference to a given class and is probably generalising its forecasts in a more or less general way. Although the general accuracy (.65) reflects average predictive power, it suggests that there are notable improvements in accuracy since a large percentage of the cases are misclassified. The fact that the macro and weighted averages have a similarity implies that imbalance on classes does not significantly affect the performance of the model when the support of the classes differs (e.g., happy: 1774 instances; disgust: 111). On the whole, the model has a moderate yet consistent performance, indicating that further optimization, such as feature representation,

Table 4.12: VGG16 + Xception Confusion Matrix on Hybrid Dataset

Emotion	Precision	Recall	F1-score
Angry	0.65	0.65	0.65
Disgust	0.65	0.65	0.65
Fear	0.67	0.63	0.65
Happy	0.69	0.66	0.67
Neutral	0.65	0.65	0.65
Sad	0.65	0.65	0.65
Surprise	0.65	0.65	0.65
Overall	0.66	0.65	0.65
Accuracy	0.65		

data addition, or model refinement to achieve more class-specific discrimination, is needed.

4.7.4 Comparative Results on HYBRID Dataset

Here, all the selective best models trained on hybrid dataset are consolidated and compared in this section. Accuracy and loss curves, confusion matrices, per-class measures and overall performance scores serve to demonstrate the effectiveness of each architecture that can be achieved in the case of FER and when it is used to detect autism spectrum disorder. Models reveal some competence in perceiving facial emotions, their accuracy, stability, and balance per-class vary considerably.

Table 4.13: Performance Comparison of Different Models on Hybrid Dataset

Model	Accuracy	F1-score	Precision	Recall
ResNet + VGG16	0.99	0.99	0.97	0.98
ResNet + Xception	0.98	0.98	0.97	0.97
Xception + VGG16	0.65	0.65	0.66	0.65

Combinations of models tend to increase their performance, with the example of ResNet+VGG16 (0.99 accuracy and F1) and ResNet+Xception (0.98 accuracy and F1) showing that hybrid architectures can make good use of complementary features and increase the robustness of classification.

The Xception+VGG16 pair has a low synergy since it has 0.65 accuracy and F1. On the whole, these findings affirm that although high-performing individual models are effective, ensemble methods can offer a remarkable enhancement of performance in case the architectures are complementary to each other. Hybrid models work best when each network brings different, complementary features. ResNet acts as a deep feature extractor, VGG16 or Xception captures detailed or spatial patterns, so these pairs improve accuracy. VGG16 + Xception perform poor and need more investigation.

Strong hierarchy of performances appeared among the reviewed models. Models based on hybrid architectures that included ResNet had the best score, with the highest score being ResNet+VGG16 (accuracy 0.99, weighted F1 0.99), followed by ResNet+Xception (accuracy 0.98, weighted F1 0.98). VGG16 and ResNet were strong smattering models (around 0.90 accuracy), but Xception by itself was mediocre (0.69). Notably, VGG16+Xception combination did not perform highly (0.65 accuracy),

which means that multimodel methods only work well, when paired architectures are complementary. ResNet residual learning offers rich and high-level semantic features effective in complementing both VGG16 and Xception yet VGG16 and Xception have overlapping feature representations, thus, adding redundancy and not providing performance improvement.

Even though the aggregate scores are robust, the analysis of errors shows that there are still some issues with the classification of disgust and fear. Such feelings include indirect facial expressions and are also underrepresented, especially in ASD-related data.

Though this problem is mitigated by the hybrid dataset over single-source datasets, the false classifications emphasize the necessity of larger and more equal ASD-oriented datasets.

In practical terms, the enhanced recognition of clinically relevant emotions, including sadness and fear, have significant implications related to the ASD-related application, including therapeutic and educational applications to help participants become more aware of their emotions and communicate socially.

Nevertheless, there are limitations such as the fairly limited size of the ASD data and the bounds of transfer learning in relation to the cultural and demographic variations. Altogether, these findings confirm the claim that hybrid dataset training boosts the FER performance and prove that architectural compatibility has a crucial role to play in multimodel systems. ResNet-based hybrids are always more effective than other combinations, whereas the VGG16-Xception combination is not effective in this task. These results indicate the clinical promise of FER systems with ASD populations and provide justification to conduct future research on larger datasets, explainability, and practice.

4.8 Chapter Summary

The applicable use and results of the 5 state of the art CNN -VGG16, resnet50, densenet121, mobile net v2, and Xception combinations For facial emotion recognition in Autism Spectrum Disorder on FER-2013, an ASD-specific dataset, and a standardized multi-dataset containing preprocessing, were inserted into the paper. Findings indicate that the hybrid dataset was better at performance, and ResNet+Xception and ResNet+VGG16 in accuracy and F1-score exhibited the highest performance, whereas the single models of VGG16 and ResNet50 did moderately and the Xception+VGG16 underperformed.

Recent obstacles still exist in identifying exaggerated emotion such as disgust and fear, owing to class disparity and breezes. The sum of the findings is that hybrid datasets and state-of-the-art CNNs are useful in ASD-specific FER, and yet larger datasets, explainability, and real-life experimentation are necessary. Overall, the findings are optimistic about hybrid dataset solutions and state-of-the-art CNN models into the progressive development of FER in ASD and the need to use larger datasets, explainability, and practical deployment testing.

Chapter 5

Conclusion and Future Work

This is the concluding chapter of the research where the implementation outcomes of Chapter 4 are synthesised. It gives a logical overview of the research findings, promotes the appellation of the work, and gives a guideline to future research, cogitating upon how the research objectives outset in Chapter 1 were met with both theoretical and empirical contributions.

The main goal of the research was to conduct a comparative study of CNN models of facial emotion recognition (FER) in Autism Spectrum Disorder (ASD). Particularly, it tested the flexibility, efficiency and performance of VGG16, ResNet50, DenseNet121, MobileNetV2, and hybrid options such as ResNet+Xception and ResNet+VGG16 when trained on transfer learning to find the most applicable architecture to the ASD-oriented FER tasks. Another important purpose was to evaluate the added value of dataset hybridisation with a larger dataset (FER-2013) and a smaller but clinically relevant dataset (Fatma Tallat ASD) by trying to balance the classes and enhance the recognition of hidden emotions, e.g. disgust and fear.

The results prove the effectiveness of hybrid networks, with ResNet+Xception and ResNet+VGG16 demonstrating the better accuracy and F1-scores, and single networks such as VGG16 and ResNet50 showing moderate performance, and others, such as Xception+VGG16, being worse.

In general, the study indicates the power of hybrid datasets and esteemed CNN based approaches to ASD-oriented FER, their possible clinical application, and future directions such as larger ASD datasets, explainable AI, and practical implementation to optimise therapeutic and social outcomes.

5.1 Summary of Findings

5.1.1 Dataset Insights

The experiments showed that dataset design is crucial for FER performance. FER-2013 provided large-scale and diverse data but lacked ASD-specific expressions. The ASD dataset by Fatma Tallat offered clinically relevant expressions but was too small, increasing the risk of overfitting. The hybrid AUTISTIC FER dataset combined the strengths of both datasets, improving recognition of ASD-related expressions and model generalisation. Overall, hybridisation helps overcome data scarcity and addresses key research gaps in ASD-focused FER.

5.1.2 Model Performance

With five CNN models, initially ResNet50, VGG16, Xception, MobileNetV2, and DenseNet121, the researchers trained them on two different datasets, FER-2013 and the ASD-specific one by Fatma Talat, to identify the most appropriate ones to use in hybrid experiments. Using the findings, the hybrid dataset was chosen to include ResNet50, VGG16, and Xception, which were developed on the basis of the two sources. The results showed a clear performance hierarchy among the models. ResNet+VGG16 performed best, achieving 99% accuracy and weighted F1-score, followed by ResNet+Xception at 98%. VGG16 and ResNet alone achieved around 90% accuracy, while Xception alone performed moderately at 69%. The VGG16+Xception combination performed poorly (65%), suggesting that combining models is only effective when their features are complementary. Overall, ResNet's strong feature extraction and residual learning made it highly effective

when combined with other architectures. The results reported above indicate the efficiency of hybrid data and well-balanced model combinations in enhancing the ability to recognize nuanced emotions, such as disgust and fear, and underscores the significance of dataset variety, hybrid models, and the possibility of more explainable AI and practical implementation in the context of ASD-related FER systems. [48].

The hybrid dataset combines different sources as generic FER13 and asd specific fatima tallat asd dataset which helps increase the reliability and diversity of the dataset and makes the model more gernalize. Building on this dataset, the proposed multi-model combines complementary feature-learning approaches to improve classification accuracy and detection performance. Thus, the combination of a reliable hybrid dataset and an accurate hybrid CNN model is the core contribution of the thesis

5.1.3 Error Analysis

The analysis of the errors provided additional feedback regarding the performance of the classes. As hypothesized, happy and surprise emotions were resolved with construct validity in all models, likely because they are more easily viewed and more represented in data sets. Fear and disgust on the other hand were challenging, and there were also frequent cases of misclassification to the adjacent levels of sad or neutral. This is also consistent with the theory of Ekman on universal versus subtle feelings, some of which are more readable and others are founded on finer details [5].

The hybrid dataset reduced these misclassification issues, though they were not completely eliminated. Although models trained on hybrid data gave better results on subtle categories compared with FER-2013-trained models, challenges remained due to poor representation of rare emotions and overlapping facial features. This highlights the importance of expanding the dataset with more diverse and balanced emotion samples. Future studies should also focus on improving data quality to further reduce misclassification and enhance overall model performance.

5.1.4 Findings for ASD-FER

The study produces significant implications on FER in ASD as it demonstrates three key outcomes. First, it emphasises the importance of dataset design by demonstrating that hybridisation of general and ASD-specific datasets is one of the ways to find the right balance between generalisability and clinical relevance. Second, it establishes the usefulness of advanced CNN models, and especially hybrid networks such as ResNet+VGG16 and ResNet+Xception, in tackling the problem of detecting subtle facial expressions related to ASD. Finally, it shows how FER systems are relevant in therapeutic and educational settings, as the accurate identification of subtle emotions like disgust and fear can aid social communications and contribute to the specific intervention. The challenges in using facial emotions are one of the essential features of ASD, according to the DSM-5 [6], and this research will be a basis of future technologies that will help a therapist, educator, and their caregivers, to give real-time feedback and tailored interventions. The contributions go further than academic contributions to the field of affective computing, which has a big potential to clinical use, and enhance social outcomes among people with ASD. The contributions are not only contributions to the academic community of affective computing but they have great implications in clinical practice.[30]

5.2 Theoretical Findings

The thesis has a number of theoretical contributions to both FER and machine learning. It empirically confirms transfer learning-trained CNNs on Autism Spectrum Disorder (ASD) showing complex models are able to respond to the atypical expressions of faces [2]. The systematic assessment of five architectures i.e VGG16, ResNet, DenseNet, MobileNet, and Xception, on ASD-specific, and hybrid datasets demonstrates that, when compared to only generic, hybrid datasets produce better results owing to the significance of both dataset design and model architecture [1], [3]. The results also validate the positive aspects of architectural

methods. Altogether, the study facilitates FER theory in non-neurotypical groups and closes gaps in the relationship of affective computing with psychology and clinical domains.

5.3 Practical Finding

Besides theoretical value, thesis gives practical information on which clinical practice, education and technology development implications can be drawn. FER systems can be used in a therapeutic session where the autistic individuals learn to identify and respond to emotions.

Since emotional recognition deficits are core features of ASD, these systems allow therapists to offer real time, adaptive feedback, which improve emotional recognition and social communication.

FER technologies can be incorporated into learning systems in educational practices to support learning in explicit ways to involve autistic students in classrooms to understand peer communication.

Technological relevance is also brought out in the study. Lightweight and efficient in design, MobileNet can be easily applied to edge devices in resource limited settings (e.g. schools and therapy centers).

Conversely, those with the best recognition performance, including Xception and DenseNet, are the best to use in back-end clinical decision-support systems where accuracy is paramount. These models show that both high-performance and lightweight solutions are feasible, depending on their needs in an application.

On the social level, FER technologies can help diminish communication barriers, improve feedback and allow autistic persons to interact more in everyday communication.

These applied additions show that FER has the potential to extract out of the laboratory research, practical applications in real-world settings that enhance social inclusion, therapeutic principles, and education of autistic people.

5.4 Limitations of the Study

The limitations of the models themselves affect the outcomes. CNNs are also strong and prone to the interpretability issue, also known as the black box problem. In case of problems with transparency and trust, clinical adoption may be delayed due to the lack of the clear explanation of the decision-making process. Also, due to the relative smallness of ASD-specific data, overfitting is a possibility because these models may develop non-generalisable spurious attributes.

It was not possible to have a very large extent of experimentation. All the models were tested and deployed in a controlled environment using Google Colab. This provided consistency but did not take into account the dynamics of real-time deployment such as latency, hard apparent variation, and noise in the environment. In addition, the variability that might have been due to differences in cross-cultural and age specificity in the display of emotions had not been addressed, which limited the application of the findings to other groups.

Finally, there must be mentioning of ethical considerations. The use of sensitive ASD data creates the problem of privacy and consent and emphasizes the need to adopt responsible data handling behaviors. Furthermore, to implement FER technologies in the clinical or educational setting, the model of responsible AI has to be applied to ensure fairness, transparency, and inclusiveness. These limitations can provide significant advice to the further researches and underline the necessity to pay an equal attention to both technical progress and ethical defence.

5.5 Future Research Directions

5.5.1 Model Explainability

Although VGG16 achieved 91% accuracy and Xception achieved 69% individually, their combination as a multimodel approach resulted in a lower accuracy of 65%. This unexpected performance degradation requires further investigation to identify

the underlying causes, such as feature incompatibility, redundancy, or ineffective feature fusion.

There is also a need to research future model architectures that are beyond the traditional convolutional neural networks. One of these promising directions seems to be the integration of hybrid CNN-transformer models i.e. the application of convolutional feature extractors and transformer-based global attention models. These architectures can extract local face information and global contextual information which could be used to extract finer emotions. Another priority is the use of explainable AI methods, such as Grad-CAM and LIME, to enhance clinical trust and transparency. It would allow therapists and educators to examine model outputs and understand more of the forecasting process. Additionally, the effectiveness of multi-task learning models addressing affective emotion of people with ASD, their direction of gaze and recognition of gestures simultaneously could offer more comprehensive information, generating more detailed descriptions of behaviour as an object of intervention creation.

5.5.2 Real-Time and Edge Deployment

To render it applicable in real life, real-time and edge deployment of FER systems are going to be needed. computation needs can be a limiting factor in resource-constrained systems. MobileNetV2 is better optimised into a low-powered version in order to have it execute on real-time in a low-power device such as a mobile phone because the speed and accessibility of the solution are the most crucial factors in a classroom or treatment room.

Moreover, pilot testing among therapists and educators should also be conducted so that, besides being prepared technically, models should also be practical and useful. To start with, through implementation of real time feedback loops wherein the systems guide the children in a series of emotion recognition exercise, the therapeutic and educational practice would be transformed, making FER a part and parcel of the daily activities.

5.5.3 Ethical and Social Dimensions

Finally, the further work should be dedicated to ethical and social aspects of FER research. Responsible data collection on vulnerable populations such as autistic children ought to be carried out using open consent procedures. The ASD datasets need to be biased as well, and it is necessary to work on biasing it since it will lead to biased predictions or unfair ones. Inclusive design processes that involve the involvement of clinicians, educators, caregivers, and ASD people with ASD per se are necessary, as it must make FER tools practical, culturally sensitive, and ethical. Integrating the concept of co-creation during systems creation will assist the researchers to see to it that the technologies are not only advancing the science field, but also the communities where they are to be utilized will be benefited.

Bibliography

- [1] World Health Organization. “Autism spectrum disorder”, 2025. URL (<http://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders>). [Online]. Accessed: Aug. 5, 2026.
- [2] J. Zeidan, E. Fombonne, A. Scora, et al. “Global prevalence of autism: A systematic review update”. *Autism Research*, 15(5):778–790, 2022. doi: 10.1002/aur.2696.
- [3] R. Sima. “What we know (and don’t know) about autism, according to science”, 2025. URL (<https://www.washingtonpost.com/wellness/2025/04/23/autism-symptoms-rates-diagnosis-treatment/>).
- [4] K. Briot, A. Pizano, M. Bouvard, and A. Amestoy. “New technologies as promising tools for assessing facial emotion expression impairments in autism spectrum disorder: A systematic review”. *Frontiers in Psychiatry*, 12:634756, 2021. doi: 10.3389/fpsy.2021.634756.
- [5] Zulfan Zainal, Melinda Melinda, and Yuwaldi Away. “Enhancing autism classification accuracy on facial images using a cnn model decision fusion”. *Journal of Image and Graphics*, 14(3):462–476, 2026. doi: 10.18178/joig.14.3.462-476.
- [6] P. Yang, R. Fang, and X. Weng. “A typical facial expression characteristics in children with autism spectrum disorder and their application in early screening”. *Advances in Psychological Science*, 33(4):588–603, 2025. doi: 10.3724/SP.J.1042.2025.0588.

- [7] S. Alsubai, A. Alqahtani, A. Alanazi, M. Sha, and A. Gumaei. “Facial emotion recognition using deep quantum and advanced transfer learning mechanism”. *Frontiers in Computational Neuroscience*, 18:1435956, October 2024. doi: 10.3389/fncom.2024.1435956.
- [8] A. R. Khan, M. Hussain, et al. “Facial emotion recognition using conventional machine learning and deep learning methods: Current achievements, analysis and remaining challenges”. *Information*, 13(6):268, 2022. doi: 10.3390/info13060268.
- [9] A. Rehman et al. “Comprehensive review and analysis on facial emotion recognition”. *Computers, Materials & Continua*, 82(1):525–565, 2025. doi: 10.32604/cmc.2025.6026.
- [10] Z. Yu and C. Zhang. “Image-Based static facial expression recognition with multiple deep network learning”. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI)*. ACM, 2015. doi: 10.1145/2818346.2830595.
- [11] A. Mollahosseini, D. Chan, and M. H. Mahoor. “Going deeper in facial expression recognition using deep neural networks”. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016. doi: 10.1109/WACV.2016.7477450.
- [12] S. Alsubai, A. Alqahtani, A. Alanazi, M. Sha, and A. Gumaei. “Facial Emotion Recognition Using Deep Quantum and Advanced Transfer Learning Mechanism”, journal = *Frontiers in Computational Neuroscience*, volume = 18, pages = 1435956, year = 2024, month = oct, doi = 10.3389/fncom.2024.1435956.
- [13] Tayebah Shamouli Javanmardi and Amir Mohammad Heydari. “A deep learning framework for classifying autism spectrum disorder from children’s facial images using a multi-scale and vit architecture and edge computing”. *Exploration of Medicine , peer-reviewed medical journal*, 6, 2025. doi: 10.37349/emed.2025.1001376.

- [14] dr Alam I. N et al. “Transfer learning technique with efficientnet for facial emotion recognition”. *Revue d’Intelligence Artificielle*, 36(4):593–600, 2022. doi: 10.18280/ria.3604xx.
- [15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “ImageNet classification with deep convolutional neural networks”. *Communications of the ACM*, 60(6): 84–90, 2017. doi: 10.1145/3065386.
- [16] C. Gao, X. Ji, Q. Zhang, C. Tu, and H. He. “FERMam: A lightweight dual-source and multi-scale fusion framework for facial expression recognition”. *Scientific Reports*, 16:13826, 2026. doi: 10.1038/s41598-026-44396-6.
- [17] Israr Ahmad, Javed Rashid, Muhammad Faheem, Arslan Akram, Nafees Ahmad Khan, and Riaz Ul Amin. “Autism spectrum disorder detection using facial images: A performance comparison of pretrained convolutional neural networks”. *Healthcare Technology Letters*, 11(4):227–239, 2024. doi: 10.1049/htl2.12073.
- [18] K. He, X. Zhang, S. Ren, and J. Sun. “Deep residual learning for image recognition”. IEEE, 2016. doi: 10.1109/CVPR.2016.90.
- [19] N. A. Al-Humaidan and M. Prince. “A classification of arab ethnicity based on face image”. *IEEE Access*, 9, 2021. doi: 10.1109/ACCESS.2021.3069023.
- [20] S. A. Albelwi. “Deep architecture based on densenet-121 model. *International Journal of Advanced Computer Science and Applications*”, 13(10):336–343, 2022. doi: 10.14569/IJACSA.2022.0131042.
- [21] F. Chollet. “Xception: Deep learning with depthwise separable convolutions”. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807. IEEE, 2017. doi: 10.1109/CVPR.2017.195.
- [22] Yonis Mehmood, A. GulzarGulzar, Qazi Mudassar Ilyas, Jabbari, et al. “SBXception: A shallower and broader xception architecture”. *Cancers*, 15(14): 3604, 2023. doi: 10.3390/cancers15143604.

- [23] R. A. Elsheikh, M. A. Mohamed, A. M. Abou-Taleb, and M. M. Ata. “Improved facial emotion recognition model based on a novel deep convolutional structure”. *Scientific Reports*, 14:29050, 2024. doi: 10.1038/s41598-024-79167-8.
- [24] A. T. Khan et al. “Plant disease detection model for edge computing devices using MobileNet”. *Frontiers in Plant Science*, 14:1253508, 2023. doi: 10.3389/fpls.2023.1253508.
- [25] M. A. Saleh, A. T. Yong, N. Marbukhari, Y. M. Yusoff, N. Mohamed, and A. Almisreb. “Facial expression recognition: A new dataset and a review of the literature”. *Turkish Online Journal of Qualitative Inquiry*, 12(6):9804–9811, 2021.
- [26] R. Yamamoto Ravenor. “AI-Based facial emotion recognition solutions for education: A study of teacher-user and other categories”. *Advances in Artificial Intelligence and Machine Learning*, 4(2):2128–2151, 2024. doi: 10.54364/aaiml.2024.42122.
- [27] T. Kanade, J. F. Cohn, and Y. Tian. “Comprehensive database for facial expression analysis”. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG 2000)*. IEEE, 2000. doi: 10.1109/AFGR.2000.840611.
- [28] F. M. Talaat. “Real-Time facial emotion recognition system among children with autism using deep learning”. *Neural Computing and Applications*, 35(18):13257–13273, 2023. doi: 10.1007/s00521-023-08430-5.
- [29] M. Pantic and L. J. M. Rothkrantz. “Automatic analysis of facial expressions: The state of the art”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000. doi: 10.1109/34.895976.
- [30] Junpeng Liu et al. Yanfeng Ji, Zhengyu Deng. “Hugging Rain Man: A novel facial action units dataset for analyzing atypical facial expressions in children with autism spectrum disorder”. *IEEE Transactions on Affective Computing*, 16(3):2287–2302, 2025. doi: 10.1109/TAFFC.2025.3558914.

- [31] K. Liddle. “Implementing the picture exchange communication system (pecs)”. *International Journal of Language & Communication Disorders*, 36 (S1):391–395, 2001. doi: 10.1080/13682820127700.
- [32] R. A. Elsheikh, M. A. Mohamed, A. M. Abou-Taleb, and M. M. Ata. “Improved facial emotion recognition model based on a novel deep convolutional structure”. *Scientific Reports*, 14:29050, 2024. doi: 10.1038/s41598-024-79167-8.
- [33] Haposan Vincentius Manalu and Achmad Pratama Rifai. “Detection of human emotions through facial expressions using hybrid convolutional neural network-recurrent neural network algorithm”. *Intelligent Systems with Applications*, 21:200339, 2024. doi: 10.1016/j.iswa.2024.200339.
- [34] Magdin. “A case study of facial emotion classification using affdex”. *Sensors*, 19(9):2140, 2019. doi: 10.3390/s19092140.
- [35] R. W. Sholikah. “Real-Time facial expression recognition to enhance emotional intelligence in autism”. *Procedia Computer Science*, 234:222–229, 2024. doi: 10.1016/j.procs.2024.03.789.
- [36] J. Kang. “Differences between low-functioning autistic children and typically developing children in facial emotion recognition”. *Journal of Clinical Neuroscience*, 81:54–60, 2020. doi: 10.1016/j.jocn.2020.09.347.
- [37] R. Brewer. “Can neurotypical individuals read autistic facial expressions? atypical production of emotional facial expressions in autism spectrum disorders”. *Autism Research*, 9(2):262–271, 2016. doi: 10.1002/aur.1508.
- [38] Petra Radočaj and Goran Martinović. “Emotion recognition in autistic children through facial expressions using advanced deep learning architectures”. *Applied Sciences*, 15(17):9555, 2025. doi: 10.3390/app15179555.
- [39] N. Cavus. “A systematic literature review on behavioral assessment of autism spectrum disorder using machine learning”. *Journal of Personalized Medicine*, 11(4):299, 2021. doi: 10.3390/jpm11040299.

- [40] M. Ruan et al. “Can micro-expressions be used as a biomarker for autism spectrum disorder?”. *Frontiers in Neuroinformatics*, 18, 2024. doi: 10.3389/fninf.2024.789.
- [41] F. M. Talaat et al. “Real-Time facial emotion recognition model based on kernel autoencoder”. *Soft Computing*, 28(9):6695–6708, 2024. doi: 10.1007/s00500-024-670.
- [42] T. Farhat et al. “Facial image-based autism detection: A comparative study”. *Computers, Materials & Continua*, 78(1):105–126, 2024.
- [43] M. Kaur and M. Kumar. “Facial emotion recognition: A comprehensive review”. *Expert Systems*, 41(10):e13670, 2024. doi: 10.1111/exsy.13670.
- [44] A. Greco et al. “Benchmarking deep networks for facial emotion recognition in the wild”. *Multimedia Tools and Applications*, 82(8):11189–11220, 2023. doi: 10.1007/s11042-022-14023-8.
- [45] Fatma M. Talaat. “Real-Time facial emotion recognition system among children with autism based on deep learning and iot”. *Neural Computing and Applications*, 35(17):12717–12728, 2023. doi: 10.1007/s00521-023-08372-9.
- [46] T. Akalya, D. Ramyachitra, M. Shabarna, and C. Legha. “Deep learning-based detection of autism spectrum disorder and emotion recognition in children”. *Pattern Recognition*, 173:112906, 2026. doi: 10.1016/j.patcog.2025.112906.
- [47] M. L. Waskom. “Seaborn: Statistical data visualization”. *Journal of Open Source Software*, 6(60):3021, 2021. doi: 10.21105/joss.03021.
- [48] S. Alsubai, A. Alqahtani, A. Alanazi, M. Sha, and A. Gumaei. “Facial emotion recognition using deep and quantum and advanced transfer learning mechanism”. *Frontiers in Computational Neuroscience*, 18:1435956, 2024. doi: 10.3389/fncom.2024.1435956.
- [49] H. He and E. A. Garcia. “Learning from imbalanced data”. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. doi: 10.1109/TKDE.2008.239.

- [50] N. V. Chawla et al. “SMOTE: Synthetic minority over-sampling technique”. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. doi: 10.1613/jair.953.
- [51] H. He et al. “ADASYN: Adaptive synthetic sampling approach for imbalanced learning”. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, pages 1322–1328. IEEE, 2008. doi: 10.1109/IJCNN.2008.4633969.
- [52] G. King and L. Zeng. “Logistic regression in rare events data”. *Political Analysis*, 9(2):137–163, 2001. doi: 10.1093/oxfordjournals.pan.a004868.
- [53] R. A. Elsheikh, M. A. Mohamed, A. M. Abou-Taleb, and M. M. Ata. “Improved facial emotion recognition model based on a novel deep convolutional structure”. *Scientific Reports*, 14(1):29050, November 2024. doi: 10.1038/s41598-024-79167-8.
- [54] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. “ImageNet: A large-scale hierarchical image database”. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. doi: 10.1109/CVPR.2009.5206848.
- [55] M. Kaur and M. Kumar. “Facial emotion recognition: A comprehensive review”. *Expert Systems*, 41(10):e13670, 2024. doi: 10.1111/exsy.13670.
- [56] Hao Li, Yiqin Luo, Tianlong Gu, and Liang Chang. “BFFN: A novel balanced feature fusion network for fair facial expression recognition”. *Engineering Applications of Artificial Intelligence*, 138:109277, 2024. doi: 10.1016/j.engappai.2024.109277.