

# **EMOTION RECOGNITION FROM POSED AND NON-POSED FACIAL EXPRESSIONS**

Thesis

Submitted in partial fulfilment of the requirements for the degree of

**DOCTOR OF PHILOSOPHY**

*by*

**RASHMI ADYAPADY R.**



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA

SURATHKAL, MANGALORE - 575 025

September 2023



## DECLARATION

*by the Ph.D. Research Scholar*

I hereby declare that the Research Thesis entitled **EMOTION RECOGNITION FROM POSED AND NON-POSED FACIAL EXPRESSIONS** which is being submitted to the **National Institute of Technology Karnataka, Surathkal** in partial fulfilment of the requirements for the award of the Degree of **Doctor of Philosophy** in Department of Computer Science and Engineering is a bonafide report of the research work carried out by me. The material contained in this Research Thesis has not been submitted to any University or Institution for the award of any degree.



Rashmi Adyapady R., 177029 177CO004

Department of Computer Science and Engineering

Place: NITK, Surathkal.

Date: September 8, 2023



## CERTIFICATE

This is to certify that the Research Thesis entitled **EMOTION RECOGNITION FROM POSED AND NON-POSED FACIAL EXPRESSIONS** submitted by **Rashmi Adyapady R.** (Register Number: 177029 177CO004) as the record of the research work carried out by her, is accepted as the Research Thesis submission in partial fulfilment of the requirements for the award of degree of **Doctor of Philosophy**.

**Dr. Annappa**  
**Professor**  
Dept. of Computer Science and Engineering  
National Institute of Technology Karnataka  
Surathkal, Post Srinivasnagar  
MANGALORE, INDIA - 575 025  
annappa@nitk.edu.in

*Annappa 11/9/23*

Prof. Annappa B.

Research Supervisor

(Signature with Date and Seal)

*Rashmi 14/09/23*

Chairman - DRPC

(Signature with Date and Seal)

Head of the Department  
Dept. of Computer Science and Engineering  
National Institute of Technology Karnataka, Surathkal  
P.O. Srinivasnagar, MANGALORE-575025



## **ACKNOWLEDGEMENTS**

My dissertation is complete, and I would like to express my gratitude. Without the help of the people who have assisted me, such advancements would not have been possible.

I want to start by sincerely thanking my research supervisor, Prof. Annappa B., for his unwavering encouragement, direction, and support during my Ph.D. journey. I thank him for identifying my potential and allowing me to work in his research team. The amount of consideration he has given during the stages of research paper and thesis writing is impeccable.

I thank Prof. P. Santhi Thilagam and Dr. Geetha V., members of my Research Progress Committee (RPAC), for their thoughtful comments and recommendations on enhancing my work. They made some insightful observations that gave me new perspectives on my work. I want to express my gratitude to the Head of the Department, Dr. Manu Basavaraju and former Head of the Department Dr. Shashidhar G. Koolagudi, Dr. Alwyn Roshan Pais and other faculty members for their continuing support. I appreciate the assistance of the entire teaching and non-teaching staff of the NITK Computer Science and Engineering Department throughout my research period.

I would like to thank all the faculty of CSE Department Dr. Soumya Hegde, Dr. Basavaraj Talwar, Dr. Mohit P. Tahiliani, Mrs. Vani M., Dr. Jenny Rajan, Dr. B. R. Chandavarkar who always supported me in all the circumstances.

A special heartfelt thanks to Mrs. Seema S. who supported and blessed me and who was always there all the time with positive wishes. I would like to thank Mr. Dayanand (Academic section) who has always supported me and many students like me all the time.

I would like to sincerely thank all the staffs of CSE Department Mrs. Vanitha Singh, Mr. Dinesh Kamath, Mrs. Yashavanthi, Mr. Vairavanathan, Mr. Pradeep D.,

Mrs. Harshitha Shetty, Mrs. Mohini, Mr. Sumedha Rao, Ms. Aishwarya S., Mr. Arun Kumar, Mr. Sukhith, Mr. Vikranth and Mrs. Supriya. Also, thanks to all the housekeeping staffs of CSE Department.

I owe all that I am or ever hope to be to my mother, Mrs. Poovamma. You are the number one problem-solver in my life. I want to thank you, without whom I could never have obtained this esteemed degree. I want to thank my father, Mr. Raghava A., and my entire family for bearing with me and supporting me during my Ph.D. journey. I want to express my sincere gratitude to my uncle Mr. Vishwanatha for helping me finish my Ph.D. work and for his constant guidance and support.

I am deeply grateful to my husband, Mr. Sachin S., for being by my side through all of my ups and down, leading me, making me strong, and making every day happy.

A heartfelt thanks to my small group of friends Shubham Dodia, Spoorthy V., Garima Pandey who were always there throughout my Ph.D. journey whenever I needed them and they have helped me in all the aspects.

Thanks to all the research scholars who have helped me in some or the other way. Special thanks to the research scholars, Dr. Srinivas Routh, Dr. Apurva Kittur, Dr. Alkha Mohan, Ms. Akhila P., Dr. Likewin Thomas, Dr. Manjunath Mulimani, Mrs. Aarabhi Putty, Dr. Pradeep Nazareth, Dr. Karthik K., Dr. Uma Priya D., Dr. Vishal Jitendrakumar Rathod, Dr. Khyamling Parane, Dr. Nagaratna B. Chittaragi, Dr. Srinivasa K., Mrs. Sadhana Shetty, Mrs. Sneha Kamble, Dr. Y. V. Srinivas Murthy, Dr. Ramteke Pravin Bhaskar, Mr. Sachin D. N., Mr. Sandeep M., Mr. Kallinatha H. D., Mr. Naveen Kumar M. R., Mr. Nitesh Naik, Dr. Ajnas Muhammad. I thank all the past and present lab-mates for the friendly and conductive atmosphere in lab.

Thank you God for all the things you have given me more than I deserve.

Rashmi Adyapady R.



# ABSTRACT

Facial Emotion Recognition is an important topic of research in the field of computer vision and artificial intelligence. It plays a vital role in analyzing the current state of user behavior through their expressions. Emotion recognition through the face is an issue that researchers in the field of affective computing have extensively addressed. This issue is usually named Facial Expression Recognition (FER). There has been considerable work done on the recognition of emotional expressions. The application of this research is beneficial in improving human-machine interaction, knowing the other person's mental state and intentions, and recognizing suspicious lies and crime detection, thus improving safety and taking prior actions in case of emergencies. Although Facial Emotion Recognition can be conducted using multiple sensors, this work focuses on facial images because visual expressions are one of the main information channels in interpersonal communication. It is challenging for machines to recognize emotions in the same way as humans do, as they vary with time, intensity, and appearance. Variations on the face, like occlusions, rotation, illumination changes, and accessories, degrade the performance of recognizing the expressions efficiently. This research presents an overview of facial expression recognition techniques based on machine and deep learning algorithms to classify posed and non-posed expressions and build an automated system to recognize engagement levels.

The first work was to select relevant features, reduce dimensionality, and detect non-posed expressions using the ensemble model. A Micro Expression Recognition (MER) system is proposed using Delaunay Triangulation (DT) and Voronoi Diagram (VD) approach to retrieve Region of Interest (ROIs) based on the Action Units (AUs) description. Finally, the extracted features are appended and fed into the ensemble of the machine learning model for classification. The combination of geometric and texture features retrieved from ROI complemented each other in getting better performance in distinguishing minute changes in facial areas. The list of observations obtained from this experiment using selected features on the micro-expression (ME) database has been

reported. The proposed MER system achieved good performance while recognizing non-posed expressions with accuracies of 76.47% and 67.19% using micro-expression databases.

Second, the task was to detect the presence of occlusions as it poses difficulty in localizing and detecting the facial region, resulting in substantial intra-expression variability. The primary goal of this work is to identify facial occlusions and minimize data loss during face recognition. Hence, the Xception network with residual attention mechanism (Xcep-RA) and Gradient-weighted Class Activation Mapping (Grad-CAM) visualization technique is proposed to localize the facial occlusions and combat erroneous predictions. The model showed accuracies of 99.85% and 98.95% on LFW-mask and RMFD datasets, respectively.

Detection of posed expressions using a deep neural network has been considered as the next objective for this thesis. Two models have been developed to recognize posed expressions with pose variations. An ensemble model with a frequency-based voting approach (FV-EffNet) and a stacking classifier approach (SC-EffNet) is adopted to deal with profile and frontal pose variations and classify the posed expressions into respective classes. The extracted features are fed into the base classifier and passed through the meta classifier to analyze the data pattern and get accurate predictions. The reason behind using a stacking classifier is that it decreases the risk of getting varied outputs from different machine learning classifiers. The proposed multi-stage posed expression classification model achieved accuracies of 98.71% and 98.56%, respectively, making the system robust against pose variations.

The assessment of engagement levels from visual cues has been considered a final objective for this thesis. The Facial Engagement Analysis-Network (FEA-Net) has been proposed for learning engagement assessment in Massive Open Online Courses (MOOC) scenarios that could help to reduce the dropout rates and overcome some of the educational problems by improving the quality of learning. In this work, the spatial and temporal features are generated by Convolutional Recurrent Neural Network (CRNN) and OpenFace features that are fused into FEA-Net, which will help discern the engaged state and improve the performance of classification of engagement levels.

The model achieved an accuracy of 62.16%. The proposed models have been evaluated on publicly available datasets, and performance is compared against state-of-the-art systems.

**Keywords:** Facial Expression Recognition, Micro-Expression Recognition, Convolutional Neural Network, Posed Expressions, Non-posed Expressions, Occlusions, Facial Engagement Recognition.



# CONTENTS

<b>List of Figures</b>	v
<b>List of Tables</b>	ix
<b>List of Abbreviations</b>	xi
<b>1 Introduction</b>	<b>1</b>
1.1 Background	1
1.2 Related Concepts	5
1.2.1 Emotion Theories	5
1.2.1.1 Basic Emotion Theory	6
1.2.1.2 Dimensional Theory of Emotion	6
1.2.1.3 Compound Emotions	7
1.2.2 Sign and Message Judgment	8
1.2.2.1 Facial Action Coding System (FACS)	8
1.2.2.2 Action Units (AUs)	9
1.2.3 Elicitation of Expressions:	12
1.2.4 Level quantization	14
1.3 Motivation	15
1.4 Challenges	15
1.5 Applications	17
1.6 Brief Overview of Thesis Contributions	18
1.6.1 Feature Extraction and Classification of Non-posed Expressions	18
1.6.2 Facial Occlusion Detection	18
1.6.3 Posed Expression Classification	18
1.6.4 Facial Engagement Analysis	19
1.7 Organization of the Thesis	19

1.8 Summary	20
<b>2 Literature Review</b>	<b>21</b>
2.1 Traditional Machine Learning (ML) Approaches used in Constrained and Unconstrained Environments	24
2.2 Deep Learning (DL) Approaches used in Constrained and Unconstrained Environments	28
2.3 Traditional Machine Learning versus Deep Learning techniques	45
2.4 Summary of FER techniques based on the Data	45
2.5 Summary of research works on FER, based on posed, non-posed expressions, facial occlusions and engagement recognition	49
2.5.1 Classification of Non-posed Expressions	49
2.5.2 Facial Occlusion Detection	51
2.5.3 Posed Expression Recognition	51
2.5.4 Facial Engagement Analysis	53
2.6 Overall Findings from the Literature Review	60
2.7 Research Gaps Identified	62
2.8 Problem Statement	64
2.9 Problem Description	64
2.10 Datasets considered for this Thesis	65
2.11 Summary	67
<b>3 Feature Extraction and Classification of Non-posed Expressions</b>	<b>69</b>
3.1 Introduction	69
3.2 Preliminaries	73
3.3 Proposed Feature Extraction and Classification Technique	75
3.3.1 Stage 1: Face Detection	77
3.3.2 Stage 2: Feature Extraction	79
3.3.3 Stage 3: Classification	87
3.4 Results and Discussions	88
3.4.1 Database Description	88
3.4.2 Experimental Setup	88

3.4.3	Comparison with state-of-the-art methods	92
3.5	Summary	94
<b>4</b>	<b>Facial Occlusion Detection</b>	<b>97</b>
4.1	Introduction	97
4.2	Preliminaries	100
4.3	Proposed model for facial occlusion detection	101
4.4	Results and Discussions	104
4.4.1	Dataset Description	104
4.4.2	Experiment Analysis and Comparisons	104
4.4.2.1	Occlusion Detection	106
4.4.2.2	Comparison with State-of-the-art Methods	108
4.5	Summary	109
<b>5</b>	<b>Classification of Posed Expressions</b>	<b>111</b>
5.1	Introduction	111
5.2	Preliminaries	114
5.2.1	EfficientNet	114
5.2.2	Stacking Classifier	115
5.3	Proposed model for classification of Posed Expressions	116
5.4	Results and Discussions	122
5.4.1	Dataset Description	122
5.4.2	Implementation Details	124
5.4.3	Experiment 1: Evaluation of EfficientNet B0 model	124
5.4.4	Experiment 2: Proposed Methodology	125
5.4.5	Observations	131
5.4.6	Experiment Analysis and Comparisons	134
5.5	Summary	136
<b>6</b>	<b>Facial Engagement Analysis</b>	<b>139</b>
6.1	Introduction	139
6.2	Proposed model for facial engagement analysis	141
6.2.1	Pre-processing	143

6.2.2	Feature Extraction	143
6.2.2.1	OpenFace Features	143
6.2.2.2	Convolutional Recurrent Neural Network (CRNN)	
	Features	144
6.2.3	Classification using Facial Engagement Analysis-Network	
	(FEA-Net)	145
6.3	Results and Discussions	145
6.3.1	Database Description	145
6.3.2	Experiment Setup and Neural Network Configuration	146
6.3.3	Comparison with state-of-the-art methods	148
6.3.3.1	Discussions	148
6.4	Summary	149
<b>7</b>	<b>Conclusions and future scope</b>	<b>151</b>
7.1	Summary and Conclusions	151
7.2	Limitations and Possible Future Directions	154
	<b>References</b>	<b>157</b>
	<b>Publications</b>	<b>175</b>



## LIST OF FIGURES

1.1	The Circumplex of Emotions (Russell [1980]).	7
1.2	The AUs corresponding to upper and lower face (De la Torre and Cohn [2011]).	10
1.3	Temporal states indicating the intensity level of the emotions expressed (Cruz et al. [2014]).	15
1.4	Outline of the tasks performed in this thesis.	19
2.1	The taxonomy of literature review representation.	21
2.2	The diversity of AUs in constrained and unconstrained environment. The Figure is adapted from Li and Deng ([2018b]).	61
3.1	Facial segmentation and feature extraction system for MER.	72
3.2	Delaunay Triangulation (DT) construction (Cheddad et al. [2008]).	74
3.3	Detailed procedure outlining the flow of proposed methodology.	76
3.4	Facial landmarks obtained from Dlib function (Shahar and Hel-Or [2019]).	78
3.5	Facial landmarks points and Delaunay Triangulation (DT).	82
3.6	Facial landmark points and Voronoi Tessellation.	83
3.7	Delaunay Triangulation (DT) based facial segmentation using Action Unit (AU) indexes as the seed.	85
3.8	Voronoi based facial segmentation using Action Unit (AU) indexes as the seed.	86
3.9	CDE and HDE result with majority voting and stacking classifier approaches.	90

3.10 Confusion Matrices (a) CASMEII→SAMM (b) SAMM→CASMEII	
(c) CASMEII→SMIC (d) SAMM→SMIC (e)	
CASMEII+SAMM→SMIC . . . . .	93
4.1 Categories of face recognition during the presence of facial occlusions.	99
4.2 The overall network structure for the localization of facial occlusions.	102
4.3 Confusion matrix results for the datasets (a) Webface-OCC (b) LFW (c)	
RMFD . . . . .	105
4.4 The localization of facial occlusions using Grad-CAM visualization	
approach. . . . .	107
5.1 Ensemble model architecture based on the frequency of votes	
(FV-EffNet). . . . .	117
5.2 Stacking classifier architecture (SC-EffNet) . . . . .	118
5.3 The five pose angles with cameras in the order 180°, 135°, 90°, 45° and 0°	123
5.4 Detailed procedure outlining the flow of proposed methodology . . . . .	126
5.5 Results obtained from base classifiers (level 0) on Oulu-CASIA dataset	126
5.6 Results obtained from base classifiers (level 0) on RaFD dataset (Multi-	
Pose) . . . . .	127
5.7 Selection of meta classifier (level 1) for evaluation of distinct set of base	
classifiers in stacking classifier approach. . . . .	129
5.8 Results obtained when evaluating the frequency-based and stacking	
classifier approaches on Oulu-CASIA dataset. . . . .	130
5.9 Results obtained when evaluating the frequency-based and stacking	
classifier approaches on RaFD dataset (Multi-Pose). . . . .	130
5.10 Confusion Matrices obtained from machine learning classifiers on	
Oulu-CASIA dataset ((a) Extra Trees Classifier (b) Random Forest	
(RF) (c) Decision Trees (DTree) (d) K-Nearest Neighbors (KNN) (e)	
Support Vector Machine (SVM) (f) Multi-Layer Perceptron (MLP)) . . . . .	131

5.11 Confusion Matrices obtained from machine learning classifiers on RaFD (Multi-Pose) dataset ((a) Extra Trees Classifier (b) Random Forest (RF) (c) Decision Trees (DTree) (d) K-Nearest Neighbors (KNN) (e) Support Vector Machine (SVM) (f) Multi-Layer Perceptron (MLP)) . . . . .	132
5.12 Confusion Matrix obtained from the combination of various machine learning classifiers on Oulu-CASIA dataset (a) FV-EffNet (b) SC-EffNet	133
5.13 Confusion Matrix obtained from the combination of various machine learning classifiers on RaFD (Multi-Pose) dataset . . . . .	133
6.1 Proposed Framework for Engagement Recognition. . . . .	142
6.2 Depthwise Separable Convolutional Network architecture (Le et al. 2021)	146
6.3 Four levels of engagement from DAiSEE Dataset . . . . .	149



## LIST OF TABLES

2.1 Dataset used in various FER tasks. ( <i>Note: Only the datasets cited in articles are listed</i> ) . . . . .	22
2.2 Summary of literature based on Traditional Machine Learning Techniques . . . . .	26
2.3 Summary of literature based on Deep Learning Techniques . . . . .	32
2.4 Summary of FER Techniques based on the Type of Data . . . . .	46
2.5 Summary of research works on non-posed expression recognition. (Note: Listed only some relevant articles considered for the comparison of our objective). . . . .	55
2.6 Summary of research works on facial occlusion detection. (Note: Listed only some relevant articles considered for the comparison of our objective). . . . .	56
2.7 Summary of research works on posed expression recognition. (Note: Listed only some relevant articles considered for the comparison of our objective). . . . .	57
2.8 Summary of research works on facial engagement recognition. (Note: Listed only some relevant articles considered for the comparison of our objective). . . . .	59
2.9 The details of datasets considered in this thesis. . . . .	67
3.1 Landmark indexes considered for evaluation of the facial feature components. . . . .	79
3.2 Mapping of action units based on landmark indexes considered for the proposed methodology. . . . .	80

3.3	Comparison of results from different machine learning classifiers on three datasets.	91
3.4	Comparison of MER performance against state-of-the-art methods evaluated on CASMEII, SAMM, SMIC(HS) datasets.	92
4.1	Experiment results obtained with the proposed methodologies on three datasets.	105
4.2	Comparison of the proposed model with previous approaches.	108
5.1	Architecture of EfficientNet B0 (Tan and Le 2019)	115
5.2	EfficientNet B0 Architecture results when utilized as a Classifier	125
5.3	Results of fine-tuning individual classifiers on Oulu-CASIA and RaFD datasets.	128
5.4	The output of distinct machine learning algorithms after fine-tuning on Oulu-CASIA and RaFD datasets.	129
5.5	Comparison with previous approaches on Oulu-CASIA and RaFD datasets	135
5.6	Experiment results using other performance metrics	136
6.1	Dataset Split for Evaluation (Abedi and Khan 2021a)	146
6.2	Comparison of the proposed model with previous approaches.	148

## LIST OF ABBREVIATIONS

<u>Abbreviations</u>	<u>Expansion</u>
300-W	300 Faces In-the-Wild
2DCNN	2 Dimensional CNN
3DCNN	3 Dimensional CNN
ACC	Accuracy
ACNN	Attention CNN
AD	Augmentation and Dropout
AFEW	Acted Facial Expressions In The Wild
AffectNet	Affect from the InterNet
AFLW	Annotated Facial Landmarks in the Wild
ANN	Artificial Neural Network
ATNet	Apex-Time Network
AU	Action Unit
AUC	Area Under the ROC Curve
AU-CNN	AU detection with CNN
AU-GACN	AU-assisted Graph Attention Convolutional Network
AU-ICGAN	AU Intensity Controllable Generative Adversarial Nets
AVEC	Audio/Visual Emotion Challenge
BAE	Binarized Auto-Encoders
BAUM-1s	Bahcesehir University Multimodal Affective Database-1
BEGAN	Boundary Equilibrium Generative Adversarial Networks
BiGRU	Bidirectional GRU
BiLSTM	Bilateral Long Short-Term Memory
BiLSTM-RNN	BiLSTM-Recurrent Neural Network
Bi-WOOF	Bilateral-Weighted Oriented Optical Flow
BNN	Binarized Neural Networks
BoVW	Bag of Visual Words
BP4D-Spontaneous	Binghamton-Pittsburgh 4D Spontaneous Expression Database
BRNN	Bidirectional Recurrent Neural Networks
BU-3DFE	Binghamton University 3D Facial Expression
C3D	Convolutional 3D
CASMEII	Chinese Academy of Sciences Micro-ExpressionII

<b>Abbreviations</b>	<b>Expansion</b>
CAS(ME) <sup>2</sup>	Chinese Academy of Sciences Macro-Expressions and Micro-Expressions
CC-CNN	Cross-Channel CNN
CDE	Cross Database Evaluation
CFP-FP	Celebrities in Frontal Profile-Frontal Profile
CGAN	Conditional Generative Adversarial Network
ChaLearn-LAP	ChaLearn-Looking at People
CHEAVD	Chinese natural Emotional Audio-Visual Database
CK	Cohn-Kanade
CK+	Extended Cohn-Kanade
CNN	Convolutional Neural Network
CoNERF	Conditional Convolutional Neural Network Enhanced Random Forest
Conv-Deconv	Convolutional Deconvolutional Networks
CPU	Central Processing Unit
CRN	Convolutional Relation Network
CRNN	Convolutional Recurrent Neural Network
CSV	Comma Separated Values
CV	Cross-Validation
CVT	Centroidal Voronoi Tesellation
D	Dynamic
DAKL	Deep Attentive Center Loss
DAiSEE	Dataset for Affective States in E-learning Environment
DAM-CNN	Deep Attentive Multipath CNN
DBM	Deep Boltzmann Machine
DBN	Deep Belief Network
DCMA-CNN	Deep Comprehensive Multi-patches Aggregation CNN
DenseNet	Densely Connected Convolutional Networks
DeRL	De-expression Residue Learning
DFEW	Dynamic Facial Expression in the Wild
DFSN	Deep Facial Sequential Network
DFSTN	Deep Facial Spatio Temporal Network
DGFN	Differential Geometric Fusion Network
DICNN	Dual Integrated CNN
Dis-ExpLet	Discriminative ExpressionLet
DISFA	Denver Intensity of Spontaneous Facial Action
DL	Deep Learning
DLP-CNN	Deep Locality Preserving-CNN
DML-Net	Dynamically Multi-channel metric Network
DNN	Deep Neural Networks
DRFS-S	Domain Regeneration in the original Feature Space with unchanged Source domain



<b>Abbreviations</b>	<b>Expansion</b>
DRFS-T	Domain Regeneration in the original Feature Space with unchanged Target domain
DRLS	Domain Regeneration in the Label Space
DSD	Dense-Sparse-Dense
DSN	Deep Spatial Network
DT	Delaunay Triangulation
DTAGN	Deep Temporal Appearance Geometry Network
DTN	Deep Temporal Network
DTree	Decision Trees
ELBPTOP	Extended Local Binary Patterns on Three Orthogonal Planes
ELM	Extreme Learning Machine
EMFACS	EMotional Facial Action Coding System
EMM	Eulerian Motion Magnification
EmotiW	Emotion Recognition in the Wild
ETI	Expressional Transformation Invariant
Extra Trees	Extremely Randomized Trees
KDEF	Karolinska Directed Emotional Faces
KNN	K-Nearest Neighbor
FABO	Face and Body Gesture
FaceVid	Face Video
FACS	Facial Action Coding System
FC	Fully Connected
FDRL	Feature Decomposition and Reconstruction Learning
FEA-Net	Facial Engagement Analysis-Network
FED-RO	Facial Expression Database with Real-world Occlusions
FER	Facial Expression Recognition
FER2013	Facial Expression Recognition Challenge 2013
FERPlus	Facial Expression Recognition Plus dataset
FG-Emotions	Fine Grained Emotions
FN	False Negative
FP	False Positive
FSR-FER	Feature level Super Resolution method for Robust Facial Expression Recognition
FV-EffNet	EfficientNet model with the Frequency based Voting strategy
gACNN	Global-local based ACNN
GB	Gigabyte
GEMEP-FERA	GEneva Multimodal Emotion Portrayals
GHZ	Gigahertz
GNB	Gaussian Naive Bayes

<b><u>Abbreviations</u></b>	<b><u>Expansion</u></b>
GPU	Graphical Processing Unit
Grad-CAM	Gradient weighted-Class Activation Mapping
GRU	Gated Recurrent Unit
HCI	Human Computer Interaction
HD	High Defination
HDE	Hold-out Database Evaluation
HOG	Histogram of Oriented Gradients
HOG-TOP	Histogram of Oriented Gradients from Three Orthogonal Planes
HS	High-Speed
ICA	Independent Component Analysis
IDSDA	Intradomain Structure Domain Adaptation
IF-GAN	Identity Free Conditional Generative Adversarial Network
IW	In-the-Wild
JAFFE	Japanese Female Facial Expression
LASSO	Least Absolute Shrinkage and Selection Operator
LBP	Local Binary Pattern
LBP-TOP	Local Binary Patterns on Three Orthogonal Planes
LDA	Linear Discriminant Analysis
LDCA	Local Discriminative Component Analysis
LED	Light Emitting Diodes
LER	Lightweight Emotion Recognition
LFFC	Light Field Face Constrained
LFFW	Light Field Faces in the Wild
LFW	Labelled Faces in the Wild
LGAttNet	Local and Global Attention Network
LGBP	Local Gabor Binary Pattern
LGBP-TOP	Local Gabor Binary Pattern on Three Orthogonal Planes
LP	Locality Preserving
LPQ	Local Phase Quantization
LRCN	Long-Term Recurrent Convolutional Network
LSTM	Long-Short Term Memory
LTS	Long-Term Support
MAX	MAXimally discriminative facial movement coding system
MBCConv	Mobile Inverted Bottleneck Conv
MDLBP	Multi-Scale Dense Local Binary Patterns
MDSTFN	Multi-channel Deep Spatial Temporal Feature Fusion Neural Network
ME	Micro-Expressions

<b>Abbreviations</b>	<b>Expansion</b>
MER	Micro Expression Recognition
MFDD	Masked Face Detection Dataset
ML	Machine Learning
MLCNN	Multi-Level CNN
MLP	Multi-Layer Perceptron
MOOC	Massive Open Online Courses
MP-Adaboost	Multi-Pose Adaptive Boosting
MPVS-Net	Multi-Path Variation Suppressing Network
MSAU-Net	Multi-Scale Action Unit based Network
MSCNN	Multi-Signal CNN
MTCNN	Multi-Task CNN
Multi-PIE	Multi Pose, Illumination, Expressions
NCM	Normalized Central Movements
NIR	Near Infrared
NME	Normalized Mean Errors
OAFR	Occlusion Aware Face Recognition
O-Net	Output Network
ORB	Oriented FAST and Rotated BRIEF
ORecFR	Occlusion Recovery based Facial Recognition
ORFE	Occlusion Robust Feature Extraction
OS	Operating System
PASM	Point Adversarial Self Mining
PatchGAN	Patch Generative Adversarial Network
PCA	Principal Component Analysis
PDLS	Prior Distribution Label Smoothing
PDSN	Pairwise Differential Siamese Network
PHOG	Pyramid Histogram of Oriented Gradient
PHRNN	Part-based Hierarchical Bidirectional Recurrent Neural Network
PLS	Partial Least Squares
P-Net	Proposal Network
PPDN	Peak Piloted Deep Network
PSR	Pyramid with Super Resolution
RAF-AU	Real-world Affective Faces Action Unit
RaFD	Radboud Faces Database
RAF-DB	Real-world Affective Faces Database
RAM	Random Access Memory
RAN	Region Attention Network
RBF	Radial Basis Function
RB-Loss	Region Biased Loss
RC	Recording Conditions
RCNN	Region based CNN

<b>Abbreviations</b>	<b>Expansion</b>
RECOLA	Remote Collaborative and Affective Interactions
ReLU	Rectified Linear Unit
RF	Random Forest
RMFD	Real-World Masked Face Dataset
RMFRD	Real-World Masked Face Recognition Dataset
RML	Ryerson Multimedia Research Lab
RMSE	Root Mean Square Error
R-Net	Refine Network
ROI	Region of Interest
S	Static
SAMM	Spontaneous Actions and Micro-Movements
SC-EffNet	EfficientNet model with the Stacking Classifier
SCN	Self-Cure Network
SE-Net	Squeeze and Excitation Networks
SERD	Salient Expressional Region Descriptor
SFEW	Static Facial Expressions in the Wild
SIFT	Scale Invariant Feature Transform
SLJDA	Subspace Learning and Joint Distribution Adaptation
SMB	Soft Mask Branch
SMFD	Simulated Masked Face Dataset
SMFRD	Synthetic Masked Face Recognition Dataset
SMIC	Spontaneous Micro-expression Corpus
SR	Super Resolution
SURF	Speeded Up Robust Features
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TB	Trunk Branch
TCN	Temporal Convolutional Network
TFD	Toronto Face Dataset
TFEID	Taiwanese Female Expression Image
TMSAU-Net	Two-stream Multi-Scale Action Unit based Network
TN	True Negative
TP	True Positive
UAR	Unweighted Average Recall
UF1	Unweighted F1-score
VD	Voronoi Diagram
VGG-Face	Visual Geometry Group-Face
VGG-Funnel	Visual Geometry Group-Funnel
VIS	Visual
W-CR-AFM	Weighted-Center Regression-Adaptive Feature Mapping
WPCA	Whitened Principle Component Analysis

**Abbreviations****Expansion**

---

WSRGB-I3D

Weighted Single RGB-stream Inflated 3D ConvNet

XceptionNet

Extreme Inception Network

Xcep-RA

Xception Network with Residual Attention



# CHAPTER 1

## INTRODUCTION

### 1.1 BACKGROUND

Human Computer Interaction (HCI) system aims at providing systematic interaction between humans and machines. Charles Darwin and Prodger has firmly placed facial expressions in an evolutionary context, and has marked the origin of a study on facial expressions (Darwin and Prodger 1998). In 1872, he first suggested that, facial expressions revealing basic emotions are universal and his ideas have been a centerpiece for the theory of evolution (Hess and Thibault 2009). Facial expression conceded to be the best way to interact or communicate one's emotions and feelings. Charles Darwin (Ekman 2009a) significant contribution was a consideration of discrete emotions; the second contribution was an emphasis on the face, as facial expression contains the most valuable source of information; and the third contribution was revealing facial expressions of emotions as universal. The fourth observation was that emotions are not unique to humans but can be seen in other species. And his fifth contribution clarified why some movements correspond to a particular emotion. Thus, this began the evolution of the theory of Facial Expression Recognition (FER). Ekman and Friesen in 1972 have proposed display rules as one of the essential aspects for the production of emotional facial expression, and interpretation of these expressions vary across cultures (Dailey et al. 2010; Russell 1991). Non-verbal components like facial expressions reveal 55% of the person's intention, verbal components, and vocal segments convey 7% and 38% of the communicated message respectively (Ghimire

## 1. Introduction

and Lee [2013; Ko [2018]]. Thus, this motivates the researchers to explore the area of FER efficiently.

Face being the most complex signal system is considered to be a highly differentiated part of the human body. The face of identical twins also differs in some aspects (Samal and Iyengar [1992]). This uniqueness of face is one of the reasons for the widespread application of FER. Psychologists have concluded that every part of the face conveys some affective information. Sometimes, it is challenging to separate the same subject's facial features in two different expressions, as they may share the same feature space (Lopes et al. [2017]). There are issues with selecting appropriate features to distinguish individuals' emotions from various categories of emotions (Zhong et al. [2014]). Expressions keep varying within the same culture (Dailey et al. [2010; Russell [1991]), and patterns may depend on environment settings, mood, and situations, making it difficult for machines to recognize them efficiently (Lopes et al. [2017]). Variations in the face, facial occlusions, head poses, and illumination also degrades the overall system's performance. A generalized approach is needed that could overcome all these variations and help in building an efficient, robust system for recognition of expressions (Feifei et al. [2018]).

Facial expression is an efficient way of emotion detection, which facilitates HCI. It is a reflection of the decision; in a social context, it initiates a social exchange or response to others. It is analyzed using Action Units (AUs) or directly considers facial emotion inferences from facial expressions (Valstar et al. [2012]). In day-to-day social life, understanding the emotional feelings of others is considered to be a fundamental component and intuition. In human communication, evaluation of facial emotions are an essential factor, which helps in providing evidence about oneself and to know the intentions of the other person (Kim et al. [2017; Li and Deng [2018a]). Facial expressions are muscle movements, whereas emotions are underlying mental states which may evoke these expressions. So, there is a difference between expressions and emotions, and they are not identical. Emotions are conscious experiences a person feels, and it involves intense mental activity. It is closely related to psychological and physiological arousal signals. In neurobiological terms, emotions are complex action



programs that are triggered by internal or external stimuli. Action programs include elements such as facial expressions, action tendencies, bodily symptoms, cognitive evaluations.

The combination of both feature extraction and classification techniques is essential for FER. The main difficulties in effectively classifying facial expressions are feature extraction and classification. The performance of the best classifier would gradually deteriorate if features were sparse (Wen et al. 2017). The handcrafted features need to be improved for extracting discriminative information, but they perform well on a small amount of training data. It is challenging to adjust these low-level features based on the input data. The deep learning models overcame these difficulties by automatically learning from raw data, representing the data on numerous levels, and containing more abstract information. The rapid development in the deep learning field has impacted various areas, including FER, and has shown promising results in identifying expressions from facial images. Despite the success, computational cost remains high, imposing difficulty in availability and accessibility.

The Micro-Expressions (MEs) are not visible to the human eye, making it challenging to capture the minute changes in the facial areas as the expressions change. As a result, automating the detection of ME is a challenging task. It is also challenging to recognize spontaneous (non-posed) expression compared to posed ones' (Sultan Zia et al. 2018). Spontaneous expressions are subtle changes in facial dynamics, characteristics, timings, facial pose variations, head movements, and illumination variations that occur frequently. Addressing such issues is of utmost importance to get a robust FER system. The techniques that work suitably fine in the posed environment may fail to generalize and improve results when a real-world environment database is considered. Spontaneous data availability is also an open research problem as most existing databases are posed or induced. In psychological research, two effective strategies like sign and message judgment, are used to measure facial expressions (Cohn et al. 2007; Martinez et al. 2017; Valstar et al. 2012). Ekman and Paul distinguished these two methods as sign-judgment and message-judgment, the details are given in Section 1.2.2. According to scientists,

facial deformations convey various efficient information for social communication (Barrett et al. 2019). Facial Action Coding System (FACS) provides better discrimination between micro-expression classes (Liong et al. 2018). Also, strong dependencies exist between the facial expressions and the AUs, which can assist in guiding the model's learning process (He et al. 2021).

Occlusions occur due to the presence of obstacles. It poses difficulty in localizing and detecting the facial region, resulting in substantial intra-expression variability caused by noise and outliers. Facial occlusions are one of the most common issues that exist in real-world images. Solving such issues is essential for improving face recognition. The occlusion obscures the face region, making it more difficult to retrieve distinctive features. Occlusions frequently occur in natural settings and are challenging in computer vision and object detection. Identification of the face is challenging in a real-world scenario as the occluded area of a face image might vary in position, size, and form. It is essential to detect occlusion as it aids in effective feature selection and an accurate face recognition process.

Facial occlusions (Min et al. 2011) occur due to the presence of obstacles like a scarf, sunglasses, hat, beards, mustaches, the appearance of hand on the face, facial hair, hair covering the frontal face. The presence of occlusions blocks the facial regions and increases the difficulty in extracting discriminative facial features, resulting in facial registration error and inaccurate face alignment (Ekenel and Stiefelhagen 2009). These factors degrade the performance of FER systems. Facial occlusions lead to high intra-expression variations due to noise and outliers (Zhang et al. 2018). And, considering these noises and outliers can be the indicators of emotions.

Head pose mainly relies on the face detection process. Yaw, roll, and pitch are the rotation angles used for estimating the orientation concerning a head-centered frame (Azmi and Yegane 2012; Li et al. 2018; Min et al. 2014). Pose variation is prone to errors and degrades the performance of the FER system. Robust estimation of the head pose leads to pose-invariant face recognition. Facial pose variation is one of the difficult tasks to be tackled (Feifei et al. 2016), and such images require some transformations like initialization and normalization for analysis. Self-occlusion is also a significant

problem that occurs due to the rigid rotation of the head and includes information loss. Hence, understanding facial occlusions and pose variations is also a key to in-the-wild FER.

One of the key subjects in educational psychology is engagement. Engagement recognition is essential for monitoring online learning for efficient learning outcomes. By monitoring the student's engagement, the teacher will acquire timely feedback, diminish the dropout rates, and overcome educational problems. Engagement recognition is influenced by the subject's emotions, movement, and other behavioral characteristics. The complex behaviors in Massive Open Online Course (MOOC) environment settings cause the automatic analysis of engagement to be challenging [Niu et al. \(2018\)](#). It is necessary to automatically evaluate MOOC participants' learning engagement to improve the quality of the learning [Liao et al. \(2021\)](#); [Shen et al. \(2021\)](#). In contrast to the classroom setting, students who are learning online can adopt any comfortable posture. So, during online learning assessment, a web camera may readily capture facial expressions, which is one of the most important aspects of learning engagement.

## 1.2 RELATED CONCEPTS

This subsection discusses specific terminologies that play a significant role in the research of FER.

### 1.2.1 Emotion Theories

Emotions consist of three components: a subjective experience, external performance, and physiological arousal ([Dong et al. 2022](#)). Facial expressions are an important tool for expressing and identifying emotions since they reflect external performance. Human emotions fall under two main categories, such as a categorical and dimensional group. The distinction is whether emotions are viewed as distinct entities or as separate dimensions.

### 1.2.1.1 Basic Emotion Theory

The term “basic” means emotions are discrete and adapts to environmental changes as it gets evolved. Basic emotion categories refer to the smallest unit, which cannot be disintegrated further into diminutive semantic labels. This little set of distinctive facial configurations are commonly associated with every human being irrespective of age, gender, culture, and socialization history. This commonality has led the researchers to consider these types of primary emotions as “universal” (Russell 1994). There are six basic emotions, such as anger, disgust, fear, happy, sadness, and surprise, as defined by Ekman and Friesen (1971). In some works of literature, a neutral state is considered along with six basic emotions as a primary emotion. Every expressed emotion is independent of each other (Posner et al. 2005) in terms of psychological, behavioral, and physiological manifestations.

Micro-Expressions (MEs) are categorized into four types of emotions: positive, negative, surprise, and others (Dong et al. 2022). Happy expression is a type of positive expression that is quite simple to induce. Positive facial expressions are much easier to recognize than negative ones, which are markedly different from anger, disgust, fear, and sadness expressions. The surprise expression is unexpected, context-dependent, and distinct from both positive and negative emotions. The “others” micro-expressions have uncertain emotional implications and can be categorized into the six fundamental emotions.

### 1.2.1.2 Dimensional Theory of Emotion

Dimensional approach divides emotions into 3-dimensional spaces like Arousal, Valence, and Dominance. The categorical emotions can be mapped into dimensional space, the circumplex model of affect introduced by Russell (1980) as shown in Figure 1.1. Arousal is used to measure the activation level of emotion, i.e., measurement of excitement, and it varies from a passive state to an active or excited state. Valence is used to measure the degree of pleasure a person is feeling, i.e., “good feeling” or “bad feeling” and this varies from state of being pleasant (positive) to unpleasant (negative). The third type of dimension is dominance, which specifies the dominant and

controlling nature of a subject. In nature, it varies from being submissive to dominant. Measuring the dominance space of emotion is complex, and hence it is often omitted, which leads to a two-dimensional approach, i.e., a valence-arousal strategy (Mollahosseini et al. 2017; Posner et al. 2005; Russell 1980).

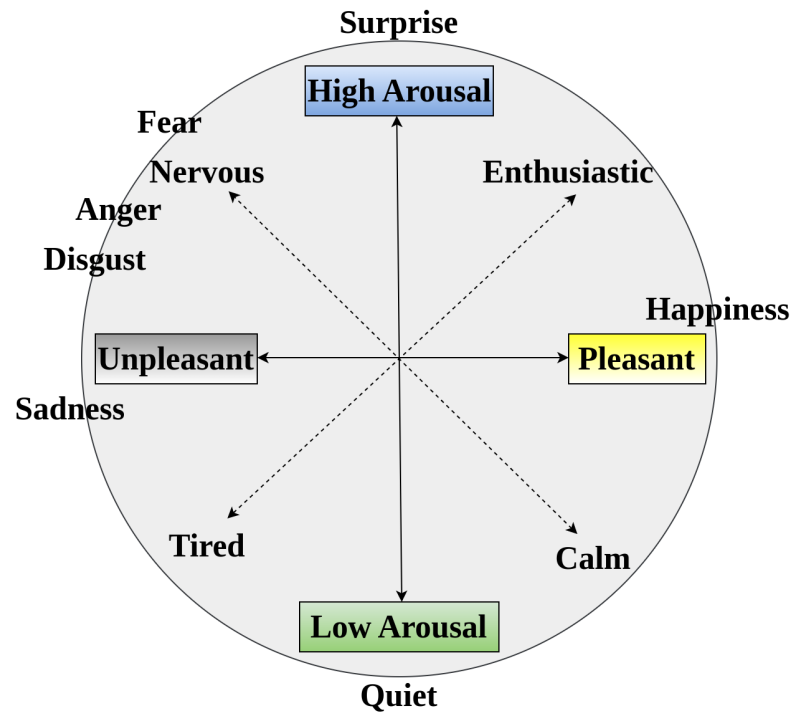


Figure 1.1: The Circumplex of Emotions (Russell 1980).

Categorical emotional states mapped onto dimensional space, a relaxed and anger state relates to low arousal space and high arousal space, respectively. While positive and negative valence relates to joy and angry state, respectively. In the case of the Pleasure-Displeasure Scale, emotions like fear and anger are unpleasant emotions, and happy is considered to be a pleasant emotion. In the Dominance-Submissiveness Scale, emotion anger is considered as a dominant emotion, while fear is considered to be a submissive emotion.

### 1.2.1.3 Compound Emotions

The term “compound” is referred to as the combination of two or more emotional feelings (Du and Martinez 2022; Du et al. 2014). The expressions such as happily-disgusted combine the facial movements considered in basic emotions happy

and disgust. Human behavior considered and displayed in compound emotions are unique and involves differential facial features. Compound emotions include dominant and complementary nature (Guo et al. 2018), and they are more detailed than the primary emotions. It is challenging to recognize compound emotions, as there exist high similarities between each variety of facial emotions. AUs of the subordinate categories are used to form the compound emotions. E.g., lip presser (AU24) and lips parted (AU25) is used to express emotions such as disgust and happy, respectively. When producing compound emotions like happily-disgusted, it is not possible to keep both AUs, lip presser (AU24) is dropped out.

### 1.2.2 Sign and Message Judgment

Sign and message judgment are two major strategies used for the measurement of facial expression in psychological research (Cohn et al. 2007; Martinez et al. 2017; Valstar et al. 2012). Ekman and Friesen (Ekman 1964) distinguished these two methods as; the observers of a sign-judgment based approach make a judgment based on “surface of behavior” shown by the subject. They count the movement of the face in a particular direction or count how long the face movement lasts or check whether the move was of frontalis or corrugator muscles (Rymarczyk et al. 2016). The observers in a sign-judgment are assumed to function like machines and called “coders”. The most widely used and well-known sign-based method is FACS. Whereas in message-judgment, observers make deductions about the “underlying facial behavior” such as affect (another term used for “emotion”) or personality and referred to as “judges” or “raters”. The raters make a judgment as angry when they see a frown expression. The coders judge the appearance as angry in terms of facial movement as having the eyebrows lowered and pulled closer together.

#### 1.2.2.1 Facial Action Coding System (FACS)

The Facial Action Coding System (FACS) is the most famous sign judgment system developed by psychologists to measure facial behavior. It helps to derive all possible facial variations in terms of Action Units (AU). AUs are caused by facial muscular contraction, and they are independent of emotions. Almost all facial expressions can

be modeled using a single AU or combination of AUs, and these AUs depict the local variations on the face. The movement of facial components encodes the maneuver of individual facial muscles to represent expression states. It includes FACS, EMotional Facial Action Coding System (EMFACS), MAXimally discriminative facial movement coding system (MAX), and AU space (Zhang et al. 2018). Ekman and Friesen in 1978 (Ekman and Friesen 1978) developed the essential and most vital approach to encode each facial behavior, which has been fine-tuned in 2002 (Chen et al. 2016; Cohn et al. 2007). The method is known as the FACS. FACS (Tian et al. 2005), (Cohn et al. 2007) is mainly used to investigate psychopathology, emotion, pain, and so on. It encodes the changes in facial muscles, in terms of AUs, which reflect discrete momentary variations in facial appearance. FACS (Sebe et al. 2005), (Martinez et al. 2017) is a standardized system for manually coding the changes in facial muscle; it has a discriminative power to characterize the actions of muscles, in terms of human emotions, by following a set of prescribed rules.

#### 1.2.2.2 Action Units (AUs)

AU is a terminology used for describing all facial actions (Yan et al. 2020). AUs are building blocks for the facial expression (Cohn et al. 2007) and are considered as the smallest visually discriminable facial movements (Lee and Ro 2016). Coders use the contraction of a single or group of muscles to create AUs. Every muscle action and its representation has a specific meaning, and each activity is assigned a unique number when producing facial expressions (Martinez et al. 2017). The FACS system is designed to detect minute changes in facial features and consists of forty-four AUs (Lee and Ro 2016). Among these, thirty AUs are related to a specific set of facial muscles corresponding to the upper and lower portions of the faces, as presented in Figure 1.2. During the evaluation, some of the AUs may occur infrequently, whereas some AUs may contribute a lot in measuring the reliability of expressions expressed (De la Torre and Cohn 2011). Investigators may pool some of the AUs to get more reliable results.































Upper Face Action Units					
AU 1	AU 2	AU 4	AU 5	AU 6	AU 7
					
Inner Brow Raiser	Outer Brow Raiser	Brow Lowerer	Upper Lid Raiser	Cheek Raiser	Lid Tightener
AU 41	AU 42	AU 43	AU 44	AU 45	AU 46
					
Lid Droop	Slit	Eyes Closed	Squint	Blink	Wink
Lower Face Action Units					
AU 9	AU 10	AU 11	AU 12	AU 13	AU 14
					
Nose Wrinkler	Upper Lip Raiser	Nasolabial Deepener	Lip Corner Puller	Cheek Puffer	Dimpler
AU 15	AU 16	AU 17	AU 18	AU 20	AU 22
					
Lip Corner Depressor	Lower Lip Depressor	Chin Raiser	Lip Puckerer	Lip Stretcher	Lip Funneler
AU 23	AU 24	AU 25	AU 26	AU 27	AU 28
					
Lip Tightener	Lip Pressor	Lips Part	Jaw Drop	Mouth Stretch	Lip Suck

Figure 1.2: The AUs corresponding to upper and lower face (De la Torre and Cohn 2011).

To recognize the emotion classes, individual AU is detected, and based on the combination of AUs; the system further classifies them into a specific category (Du et al. 2014; Tian et al. 2005). The combination of a few AUs effectively helps in representing subtle facial changes and a large variety of expression states. However, it is challenging and tedious to accurately detect and track every AU information in images because of minute changes in the facial muscles. It is challenging for psychologists to define every facial expression with the definite prototypical AUs and translate these emotion-related AUs into affective states. Differences in culture, the requirement of high-quality video equipment, high resource intense, arduous manual coding, expensive nature, and also the way of perceiving the facial expression hinders the progress towards fully automated or computer-based facial expression analysis. Expertise takes months together to learn and be professional in AU coding. The



EMFACS focuses only on those facial actions which are likely to have emotional significance in it (Yan et al. 2020). The MAX system codes discrete emotional states based on a set of formulas obtained via facial movements (Rinn 1984). It discriminates various facial movements in each facial region and even distinguishes facial movements in three facial areas, forehead and eyebrows, midface (eyes and cheeks), and mouth (Matias et al. 1989). AU space-based emotion representation uses continuous coordinates rather than binary values. It is flexible and reduces the misclassification as it is more tolerant towards the AU detection results near marginal areas (Zhao et al. 2015).

With facial AU coding, one can get the knowledge of separating three facial expression categories: Macro, Micro, and Subtle expressions.

- Ordinary facial expressions or macro-expressions cannot always reflect a person's true emotions. Macro-expressions occur over individual or multiple regions of the face depending on the expressed expressions, and it can be observed easily in daily interactions (Shreve et al. 2011), (Qu et al. 2017). Such expressions occur between 2 to 3 seconds in duration (Liong et al. 2018) and involve the entire face.
- Micro-expressions (MEs) occur when a person is trying to conceal or repress the felt emotional state consciously or unconsciously, and it occurs in a small region of the face (Ekman and Friesen 1969; Polikovsky et al. 2009; Qu et al. 2017). These expressions are very rapid and involuntary and give a brief glimpse of undergoing feelings of a person which he/she is trying to conceal (Ekman 2009b; Pfister et al. 2011). It lasts between 1/5 to 1/25 second duration in precise length (Liong et al. 2018). Subtle changes, in facial appearance, visual differences between human beings, and less number of frames, make it difficult for analysis of ME. MEs are further categorized into three categories, simulated, neutralized, and masked expressions (Bhushan 2015).

### 1.2.3 Elicitation of Expressions:

Expressions can be elicited and collected in multiple ways, such as posed, spontaneous, and in-the-wild.

- In posed appearance, subjects deliberately display the expressions by reproducing specific deformations in the facial muscle; these expressions are elicited based on the guidance of professionals or actors. Subjects elicit series of expressions based on the demand of instructors; basically, they are aware of being recorded. Free production, ordered production, and portrayal are the three ways of reproducing the posed expressions from the subjects (Weber et al. 2018).
- Spontaneous (non-posed/authentic) expressions occur naturally and are not controlled by the subjects (Mavadati et al. 2013; Sebe et al. 2007). It occurs when the subjects try to express internal feelings. Not all individuals express facial expressions in the same way; it depends on one's culture, personal, and familial display rules. Two ways of eliciting such expressions depend on passive and active approaches. In a passive approach, specific emotional states are induced by displaying images or videos, or another way is recording during the interaction of two protagonists to obtain emotion-rich content (Weber et al. 2018). In the active approach, capturing of real emotions is done directly, involving the participants themselves. Spontaneous expressions are distinct from posed ones in terms of spatial patterns, temporal patterns, morphological and dynamic properties (Namba et al. 2017; Wang et al. 2015).
- Data obtained from an unconstrained (real-time) environment includes complex emotions and variations like head pose, occlusions, illumination variation, rotations, and referred to as in-the-wild expressions. It is challenging to recognize facial expressions from these types of datasets. Three modes of getting in-the-wild emotions are crowdsourcing, the corpus of videos, or images obtained using web crawling (both posed and spontaneous expressions) (Weber et al. 2018).

Five factors that need to be considered when eliciting emotions (Picard et al. 2001) are:

1. Subject-elicited versus event-elicited: whether the expressed emotion was evoked forcefully or using external stimuli.
2. Constrained versus unconstrained settings: whether data recording was done in a laboratory or unconstrained real-world environment settings.
3. Expressed expression versus feeling: whether the emphasis was placed on externally expressed emotion or felt internal feelings.
4. Open-recording versus hidden-recording: whether the participants participating in the experiment were aware of being recorded or not.
5. Emotion-purpose versus other-purpose: whether the participant is aware of an experiment and knows that he/she is part of an analysis.

#### **Discussions on the pros and cons of each elicitation approaches.**

- a. Posed Expressions: This type of expression have different activations of facial muscles and dynamics. People tend to control and move their facial muscles, intentionally (Mavadati et al. 2016). The data collected from this environment will be highly controlled, i.e., frontal exaggerated expressive faces with minimal illumination variations and occlusions (Liu et al. 2020). These data cannot reflect real circumstances and it is hard to generalize well in real applications.
- b. Spontaneous Expressions: This type of unconscious expression links to the emotional state of an individual. AU's play an essential role in analyzing and describing facial behavior (Nonis et al. 2019). Developing an intelligent HCI system capable of understanding humans' real expressions is crucial as it will be useful to deploy in real-world applications (Mavadati et al. 2016). The dynamics of facial actions are problematic for FACS coders to simultaneously measure multiple AUs' intensity. Thus, it makes an annotation of such a database a tricky process.
- c. In-the-Wild Expressions: Such realistic facial data plays an essential role in advancing research on facial expression analysis systems (Dhall et al. 2011).

This type of expression includes unconstrained facial expressions, varied head poses, occlusions, and illuminations (Dhall et al. 2014). Compared to lab-controlled datasets, in-the-wild datasets impose a challenge making feature extraction a tedious task due to interference (Liu et al. 2020).

### **Discussions on neurological aspects for recognizing emotion from facial expressions.**

A large number of psychological studies have focused on the recognition of emotions from facial expressions over several decades (Adolphs 2002). Muscle actions are the salient features of facial expression. Facial muscles are not only the ones that respond to emotion; even striated muscles in the neck, back, arms, and also smooth muscles of the blood vessels and alimentary tract are also responsible. Annotators can objectively measure the facial expressions without knowing the semantic meaning of emotion expressed (Rinn 1984), by analyzing the position and movement of facial skin and fascia causing wrinkles, lines, folds, and facial landmarks. To describe facial actions, muscles are not directly visible. Thus, FACS is the most intricate instrument used to translate skin movements into muscle patterns. The reduction of the facial expressions into the list of AUs has the advantage, of providing the means of describing any facial configuration even when it does not willingly fit into a preconceived category.

#### **1.2.4 Level quantization**

Level quantization has four temporal states (Liong et al. 2018; Posner et al. 2005) like neutral, onset, apex, and offset phases as depicted in the Figure 1.3. These temporal segments indicate the intensity level of the emotions expressed by a person (Cruz et al. 2014). Most of the expressions start from a neutral state. A neutral phase is an expressionless phase, where no signs of muscular contraction or activities are visible. Onset state denotes the beginning state of emotion, i.e., when the facial muscle starts to contract and increases in intensity. In the apex phase, the muscular contraction is at the peak, and the depth of expression reaches a firm level. The offset phase is when the relaxation of muscular action takes place, and the intensity level reaches the saturation state. Usually, the intensity of expressions keep varying in order, such as neutral-onset-apex-offset, and there are possibilities of multiple apex states for all the

classes of emotions as well.

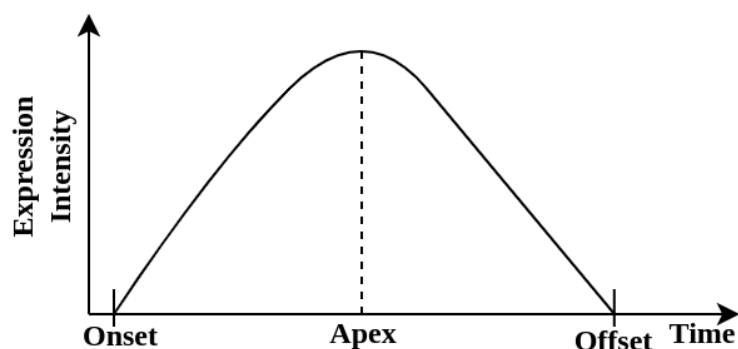


Figure 1.3: Temporal states indicating the intensity level of the emotions expressed (Cruz et al. 2014).

### 1.3 MOTIVATION

Emotions are an essential means of communication to know ones' internal mental status and real feelings. Emotions keep varying across people with respect to time and context. The variations in the intensity of the emotion expressed by the same person for the same expression would be a challenging factor for accurate recognition of emotions. Due to the variation in facial visual aspects, there could be chances of false predictions. Due to several changes observed across the human faces, the process of detecting and recognizing the emotions has become a difficult task for computers. Analyzing posed and spontaneous (non-posed) expressions to identify natural or fake emotions would be helpful for various applications such as lie detection, monitoring the facial expressions of patients suffering from disorders, students engagement etc. Further, a localization mechanism is needed to address the issues of facial occlusions. Hence, the development of an automated system for the classification of facial expressions and engagement analysis is required.

### 1.4 CHALLENGES

- Learning semantic information for identifying the subtle difference between expressions is challenging. Also, the selection of features to distinguish each facial expression from the rest is a crucial task.

- It is challenging to separate the same subject's facial features in two different expressions, as they may share the same feature space. There are issues with selecting appropriate features to distinguish individuals' emotions from various categories of emotions.
- It is challenging to recognize spontaneous expression compared to posed ones'. Spontaneous expressions are subtle changes in facial dynamics, characteristics, timings, facial pose variations, head movements, and illumination variations that occur frequently.
- The techniques that work suitably fine in the posed environment may fail to generalize and improve results when a real-time scenario is considered where a real-world environment database is used for evaluation. Thus, evaluating real-world databases is challenging compared to lab-controlled databases.
- Recognizing emotion classes like anger, sadness, fear, and disgust in both constraint and unconstrained environments is challenging. Confusion arises when classifying these universal expressions.
- Building an automated framework for Micro-Expression Recognition (MER) is difficult and needs further research.
- The presence of occlusions blocks the facial regions and increases the difficulty in extracting discriminative facial features, resulting in facial registration errors and inaccurate face alignment. Facial occlusions lead to high intra-expression variations due to noise and outliers. These noises and outliers can be indicators of emotions.
- Pose variation is prone to errors and degrades the performance of the FER system. Robust estimation of the head pose leads to pose-invariant face recognition. Self-occlusion is also a significant problem that occurs due to the rigid rotation of the head and includes information loss. Hence, understanding facial occlusions and pose variations is also a key to in-the-wild FER.
- Visual privacy-preserving is challenging. Reliable and accurate privacy-preserving methods are important in human-machine conversation and automatic FER system.

- Group-level emotion recognition from images is challenging and requires further exploration as it includes low-resolution images with cluttered backgrounds. Additionally, the recognition rate decreases, and it becomes harder to distinguish between two or more faces in a single picture sequence.

## 1.5 APPLICATIONS

This section describes various applications of analyzing the facial expressions<sup>1</sup>. In the field of marketing, recording facial expressions adds quantitative data to self-reports about a product or service. Based on the study of facial expressions, market segments can be measured, and goods can be optimised. In the media and advertising industry, the audience's emotional reaction, as seen on their faces, aids in recognising movie scenes (Navarathna et al. 2017) and rating them as positive or negative, thereby increasing positive emotions during the final release. Monitoring the facial expressions of patients (Edla et al. 2018) suffering from disorders can significantly promote the success of the underlying cognitive-behavioral therapy, both during the diagnosis and intervention phase in psychological research and medical applications. In website design, monitoring facial expressions of users while handling software or navigating websites helps provide insights such as satisfaction or dissatisfaction and, in turn, gain benefit. The process of automating and understanding how instructors judge students' engagement via face is an essential application in educational research (Whitehill et al. 2014). Engagement recognition helps the instructors improve the teaching strategy and improve instructional videos based on the viewers' engagement signals. Recognition of facial expression via surveillance camera helps in lie detection and suspicious crime detection. Such crucial information can benefit crime agencies to improve the safeness and take prior actions in case of emergencies. Physical fatigue recognition from facial expression (Lee et al. 2018), in turn, helps to alert the drivers for safeness. There is a need to develop a robust system to cope with these applications, as it involves recognition of facial expressions from unconstrained environments.

---

<sup>1</sup><https://imotions.com/blog/facial-expression-analysis/>

## **1.6 BRIEF OVERVIEW OF THESIS CONTRIBUTIONS**

The significant contributions of this thesis include the challenges in FER, like the classification of posed and spontaneous expressions. This work introduces techniques that recognize the posed expressions with multi-pose variations, classify the spontaneous Micro-Expressions (MEs), localize the facial occlusions, and analyze engagement levels in Massive Open Online Courses (MOOC) scenarios. A brief summary of the work is provided below.

### **1.6.1 Feature Extraction and Classification of Non-posed Expressions**

The task is to extract meaningful features to capture the minute changes that occur on the face and classify these non-posed micro-movements into respective classes. The geometric and texture features are obtained from the Region of Interest (ROI). The combination of geometric and texture features is fed into the ensemble models for classification. The ensemble model and the proposed technique demonstrated their effectiveness in identifying the non-posed expressions. The Cross-Database Evaluation (CDE) and Hold-out Database Evaluation (HDE) performed on the spontaneous ME datasets proved the robustness of the proposed model and thus can be helpful for real-time processing.

### **1.6.2 Facial Occlusion Detection**

Occlusions occur due to the presence of obstacles; it poses difficulty in localizing and detecting the facial region, resulting in substantial intra-expression variability caused by noise and outliers. In this work, the task is to identify whether the face is occluded or not. The Xception Network and a residual attention (Xcep-RA) method assisted in the detection of the occluded facial regions. The proposed Xcep-RA achieved a competitive result on in-the-wild datasets.

### **1.6.3 Posed Expression Classification**

The task is to extract relevant features and classify the posed expressions. In this work, the intermediate features are extracted from the deep learning models and fed to ensemble models to improve the system's performance. The proposed approach



achieved better performance on both single and multi-pose datasets, making the model robust against pose variations.

### 1.6.4 Facial Engagement Analysis

Engagement recognition is essential for monitoring online learning for efficient learning outcomes. The proposed Facial Engagement Analysis-Network (FEA-Net) helps capture valuable features for classifying engagement levels. The experiments showed comparative results on the Dataset for Affective States in E-learning Environment (DAiSEE) dataset.

## 1.7 ORGANIZATION OF THE THESIS

The thesis advances in 7 chapters as depicted in Figure 1.4. An outline of each chapter is given below.

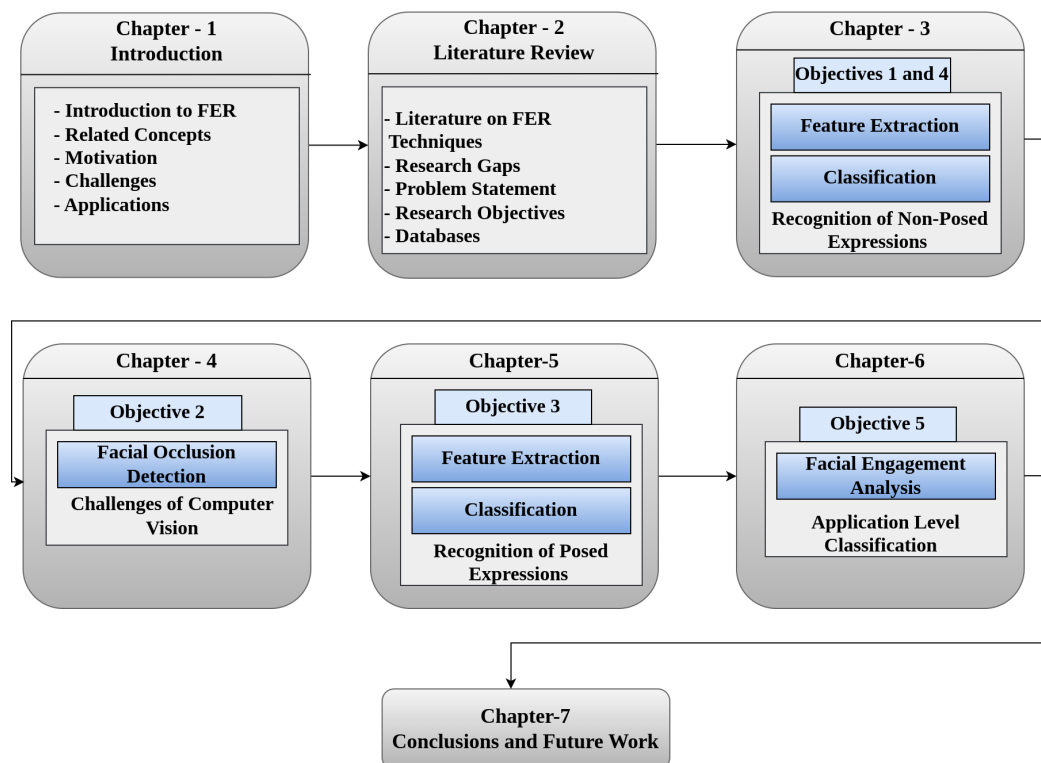


Figure 1.4: Outline of the tasks performed in this thesis.

- **Chapter 1 : The Introduction** section provides an overview of Facial Expression Recognition (FER) and detail of its related concepts. The chapter ends with a

brief overview of research contributions and a thesis outline.

- **Chapter 2 : Literature Review** section mainly aims to give deeper insight into machine learning and deep learning techniques used in both constrained and unconstrained environments, with the discussion on shortcomings and prospective future directions.
- **Chapter 3 : Feature Extraction and Classification of Non-posed Expressions** covers extraction of features from segmented facial regions and classify non-posed expressions using ensemble of machine learning classifiers.
- **Chapter 4 : Facial Occlusion Detection** includes the task of identifying the occlusions from the occluded facial images using deep learning architectures.
- **Chapter 5 : Posed Expression Classification** discusses the categorization of posed expressions using deep learning and ensemble of machine learning classifiers.
- **Chapter 6 : Facial Engagement Analysis** discusses further the analysis of engagement levels from facial images using deep learning architectures.
- **Chapter 7 : Conclusions and Future Scope** chapter summarize the contributions and findings of this research work with future scope.

## 1.8 SUMMARY

This chapter presented an overview of Facial Expression Recognition (FER) and detailed its related concepts, along with a discussion on motivation and challenges in the area of FER with its applications. Chapter 2 aims to give a deeper insight into machine learning and deep learning techniques used in both constrained and unconstrained environments. A list of research gaps identified during the critical literature analysis and the problem statement of the current work is also provided. At the end of the chapter, details of the datasets utilized to carry out the experiments are also highlighted.

## CHAPTER 2

### LITERATURE REVIEW

In chapter 1, an overview of FER and detail of its related concepts were provided, along with the discussion on motivation. This chapter aims to give deeper insight into machine learning and deep learning techniques used in both constrained and unconstrained environments, with the discussion on shortcomings and prospective future directions that may help the researchers and newcomers better comprehend the opportunities in the area of FER. This chapter also includes sections on the datasets used for the experimentation of FER.

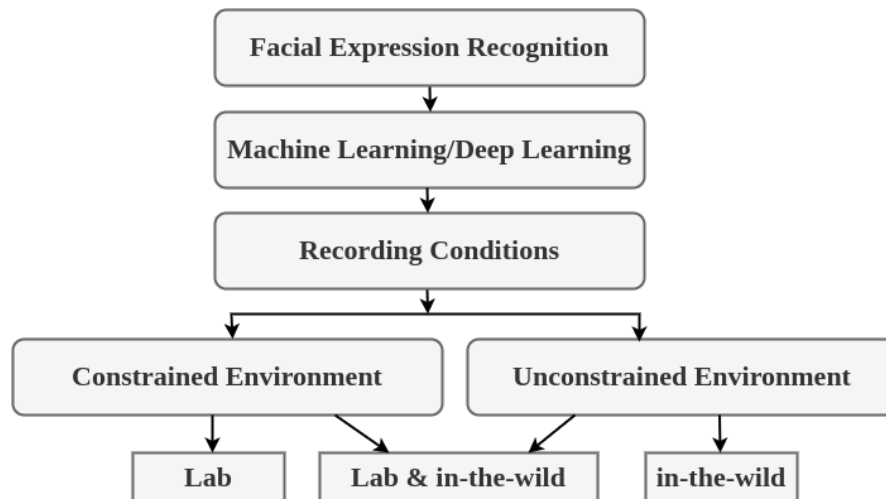


Figure 2.1: The taxonomy of literature review representation.

The chapter also presents the literature on leading traditional and Deep Learning (DL) techniques for FER. As shown in Figure 2.1, the literature review is further subdivided based on recording conditions, i.e., whether the dataset is recorded in controlled

## 2. Literature Review

---

(constrained) or/and uncontrolled (unconstrained) environment situations. In this work, if the dataset is recorded in a constrained environment, it is classified under the lab recording condition. And, if the dataset is recorded in an unconstrained environment, the recording condition is In-the-wild, termed as IW, in the following sections. If the authors have utilized the datasets recorded in constrained and unconstrained environments, the recording conditions are filled with lab and IW. Table 2.1 highlights a list of datasets used in various FER tasks.

Table 2.1: Dataset used in various FER tasks. (*Note: Only the datasets cited in articles are listed*)

Dataset	P	SP	P & SP	IW
Extended Cohn-Kanade (CK+)			✓	
Cohn-Kanade (CK)	✓			
MMI Facial Expression Database (MMI)			✓	
GENeva Multimodal Emotion Portrayals (GEMEP)-FERA		✓		
Chinese Academy of Sciences Macro-Expressions and Micro-Expressions (CAS(ME) <sup>2</sup> )		✓		
Chinese Academy of Sciences Micro-Expression dataset (CASME II)		✓		
Spontaneous Micro-expression Corpus (SMIC) [High Speed (HS), Near Infra-Red (NIR), Visual (VIS)]		✓		
Japanese Female Facial Expression (JAFPE)	✓			
Spontaneous Actions and Micro-Movements (SAMM)		✓		
Audio/Visual Emotion Challenge (AVEC)				✓
Acted Facial Expressions In The Wild (AFEW)				✓
Facial Expression Recognition Challenge 2013 (FER 2013)				✓
Binghamton University 3D Facial Expression (BU-3DFE)	✓			
Multi Pose, Illumination, Expressions (Multi-PIE)	✓			
Oulu-CASIA	✓			
Face and Body Gesture (FABO)	✓			
Ryerson Multimedia Research Lab (RML) Emotion Database	✓			
eINTERFACE05		✓		
Bahcesehir University Multimodal Affective Database-1 (BAUM-1s)		✓		
Radboud Faces Database (RaFD)	✓			
Denver Intensity of Spontaneous Facial Action (DISFA)		✓		

\* P–Posed, SP–Spontaneous, IW–In-the-Wild

Dataset	P	SP	P & SP	IW
Binghamton–Pittsburgh 4D Spontaneous Expression Database (BP4D-Spontaneous)		✓		
Toronto Face Dataset (TFD)			✓	
Emotion Recognition in the Wild (EmotiW)				✓
ChaLearn-Looking at People (ChaLearn-LAP)				✓
Remote Collaborative and Affective Interactions (RECOLA)		✓		
Real-world Affective Faces Database (RAF-DB)				✓
Static Facial Expressions in the Wild (SFEW)				✓
Labeled Faces in the Wild (LFW)				✓
Facial Expression Dataset with Real-world Occlusions (FED-RO)				✓
Affect from the InterNet (AffectNet)				✓
Taiwanese Female Expression Image (TFEID)	✓			
Face Expression Recognition Plus dataset (FERPlus)				✓
Karolinska Directed Emotional Faces (KDEF)	✓			
Dynamic Facial Expression in-the-Wild (DFEW)				✓
Chinese natural Emotional Audio–Visual Database (CHEAVD)				✓
Face Video (FaceVid)				✓
Real-world Affective Faces Action Unit (RAF-AU)				✓
WebEmotion				✓
EmotioNet 2020				✓
Fine Grained Emotions (FG-Emotions)				✓
FERFIN				✓
Light Field Faces in the Wild (LFFW)				✓
Light Field Face Constrained (LFFC)				✓
300 Faces In-the-Wild (300-W)				✓
Annotated Facial Landmarks in the Wild (AFLW)				✓
Dataset for Affective States in E-Environments (DAiSEE)		✓		
AR face database	✓			
Webface-OCC				✓
Celebrities in Frontal Profile-Frontal Profile (CFP-FP)				✓
AgeDB-30				✓
Real-World Masked Face Recognition Dataset (RMFRD)				✓
Synthetic Masked Face Recognition Dataset (SMFRD)				✓
Real-World Masked Face Dataset (RMFD)				✓
Simulated Masked Face Dataset (SMFD)				✓
MegaFace Challenge				✓
Masked Face Detection Dataset (MFDD)				✓

\* P–Posed, SP–Spontaneous, IW–In-the-Wild

## 2.1 TRADITIONAL MACHINE LEARNING (ML) APPROACHES USED IN CONSTRAINED AND UNCONSTRAINED ENVIRONMENTS

Handcrafted features are used to classify the expressions into respective emotion classes, and Table 2.2 highlights a few works of literature based on traditional machine learning techniques. To recognize the expressions collected in a constrained and unconstrained environment, most of the past works in literature have utilized various geometric and texture feature extraction techniques to extract features and fed into traditional classifiers for classification. Techniques such as Local Binary Patterns (LBP) and its variants, Histogram of Oriented Gradients (HOG) and its variants, Oriented FAST and Rotated BRIEF (ORB) feature descriptors, Bi-Weighted Oriented Optical Flow (Bi-WOOF), gabor filters have been utilized to extract features from local and global regions from both static and dynamic type of data. Further, those features are fed into traditional classifiers like Support Vector Machine (SVM) for classification.

To recognize the expressions collected in a constrained environment, Ghimire et al. (2017) has concatenated geometric feature descriptor Normalized Central Moments (NCM) with LBP and fed it to SVM for classification. The results showed that feature descriptors extracted from domain-specific local regions outperform holistic representations. Likewise, Zhong et al. (2014) has utilized LBP and its variant uniform LBP to extract features and has explored general and specific information about different facial expressions using a two-stage Multi-Task Sparse Learning (MTSL). Similarly, Niu et al. (2021) had used LBP and improvised oriented FAST and rotated BRIEF (ORB) to extract discriminative features and SVM for classification. The improvised ORB solved the problem of conventional ORB by using region-wise division for feature point extraction. Liang et al. (2018) has utilized LBP from Three Orthogonal Planes (LBP-TOP) features and Bi-Weighted Oriented Optical Flow (Bi-WOOF) feature extractor to encode crucial expression present at the apex frame of video sequences. Boughida et al. (2022) had utilized Gabor filters to extract features from Region of Interest (ROI); Principal Component Analysis (PCA) was used to choose the best feature, and a genetic algorithm was employed to optimize SVM

## *2.1. Traditional Machine Learning (ML) Approaches used in Constrained and Unconstrained Environments*

---

hyperparameters for FER. To detect minute changes in the facial region, [Guo et al. \(2019\)](#) proposed Extended Local Binary Patterns on Three Orthogonal Planes (ELBPTOP). The authors explored the second-order discriminative information in two directions of a local patch, and one is the radial differences (RDLBPTOP), and the other one is the angular differences (ADLBPTOP), as a complement to the differences between a pixel and its neighbors (LBPTOP).

Similarly, in an unconstrained environment, [Cruz et al. \(2014\)](#) has utilized LBP and its variant uniform LBP to extract features, and [Chen et al. \(2016\)](#) has used both visual and audio modalities, and obtained features using a Histogram of Oriented Gradients from TOP (HOG-TOP).

Table 2.2: Summary of literature based on Traditional Machine Learning Techniques

RC	Reference	Dataset	Type of Data	Approaches	Remarks
Lab	Ghimire et al. (2017)	CK+	S	LBP, Normalized Central Moments (NCM), SVM	<ul style="list-style-type: none"> <li>Incremental search approach has been used to determine important local regions, which has reduced the dimensions of feature and improve recognition accuracy.</li> <li>FER from local region selection has reduced the computation complexity of the algorithm.</li> <li>The performance of the proposed system has decreased when neutral emotion class has been included for evaluation as it is confused with anger and sadness emotion classes.</li> <li>Performance could be enhanced to discriminate facial expressions by searching and selecting the best features within the framework.</li> </ul>
	Zhong et al. (2014)	CK, MMI, GEMEP-FERA	D	Uniform LBP, SVM	<ul style="list-style-type: none"> <li>The proposed framework improved recognition accuracy by combining common and specific facial patches at different scales. It provides more accurate appearance locations.</li> <li>The recognition rate of anger emotion class has not been as good as when compared to other emotion classes.</li> <li>Head pose variation and facial occlusions have not been considered in this work.</li> </ul>
	Liong et al. (2018)	CAS(ME) <sup>2</sup> , CASME II, SMIC-HS, SMIC-NIR, SMIC-VIS	S	Local Binary Patterns on Three Orthogonal Planes (LBP-TOP), Bi-WOOF, SVM	<ul style="list-style-type: none"> <li>Reduces computational complexity as well as cost by selecting only apex frames instead of the entire video sequence.</li> <li>Justification is needed to understand the extent, these apex frames influence the performance of recognition.</li> </ul>
	Niu et al. (2021)	CK+, JAFFE, MMI	S & D	LBP, ORB, SVM	<ul style="list-style-type: none"> <li>Overcame excessive hardware specification requirement issues of deep learning models.</li> <li>Improved ORB solved the problem of redundancy and feature point overlap in extraction process.</li> <li>On the CK+ database, the proposed solution had a low recognition rate. Requires further research to improve accuracy and computation speed.</li> </ul>

\* RC–Recording Conditions, IW–In-the-Wild, S–Static, D–Dynamic



RC	Reference	Dataset	Type of Data	Approaches	Remarks
Lab	Boughida et al. (2022)	JAFFE, CK, CK+	S	Gabor filter, Principal Component Analysis (PCA), Genetic Algorithm	<ul style="list-style-type: none"> <li>For datasets with multiple features, the proposed technique slows down genetic algorithm convergence.</li> <li>Gabor filter parameters are manually selected to fetch the best features, but this does not guarantee the optimal values, necessitating further investigation.</li> </ul>
	Guo et al. (2019)	SMIC (HS), CASME II, SAMM	D	Extended Local Binary Patterns on Three Orthogonal Planes (ELBPTOP)	<ul style="list-style-type: none"> <li>Introduction of Whitened Principal Component Analysis (WPCA) to Micro Expression Recognition (MER) obtained more compact and discriminative feature representations, thus achieving computational savings.</li> <li>The proposed approach preserves gray scale invariance.</li> </ul>
	Jia et al. (2018)	CK+, CASMEII	S	LBP, LBP-TOP, Group Least Absolute Shrinkage and Selection Operator (LASSO), Singular Value Decomposition (SVD)	<ul style="list-style-type: none"> <li>Relation between macro and micro expressions was established to improve the MER accuracy.</li> <li>Group LASSO and SVD helped overcome the dimensionality reduction issue by selecting the most important patches and features.</li> <li>The limitation of labelled trained data was overcome by the macro-to-micro transformation approach.</li> </ul>
Lab & IW	Cruz et al. (2014)	CK, MMI, AVEC 2011 & 2012	D	LBP, SVM	<ul style="list-style-type: none"> <li>The proposed method reduces memory cost by downsampling the number of frames in video samples.</li> <li>The frames have been segmented in an evenly-sized manner and may cause a boundary effect if the unlabeled apex is spotted near to the segmentation boundary.</li> </ul>
	Chen et al. (2016)	CK+, GEMEP-FERA2011, AFEW 4.0	D	Histogram of Oriented Gradients from Three Orthogonal Planes (HOG-TOP), geometric warp feature, multiple kernel SVM	<ul style="list-style-type: none"> <li>HOG-TOP is compact and efficient in characterizing dynamic changes.</li> <li>The geometric warp feature has been useful in capturing facial configuration changes.</li> <li>The performance of in-the-wild datasets is lower compared to a lab-controlled environment, and requires further exploration.</li> </ul>

2.1. Traditional Machine Learning (ML) Approaches used in Constrained and Unconstrained Environments

\* RC–Recording Conditions, IW–In-the-Wild, S–Static, D–Dynamic

## 2.2 DEEP LEARNING (DL) APPROACHES USED IN CONSTRAINED AND UNCONSTRAINED ENVIRONMENTS

Deep learning research has tremendous success in many computer vision applications in recent years. Deep learning methods stack a number of intermediate layers from input data to a classification layer and can learn high-level semantic features automatically from a large amount of training data. Convolutional Neural Network (CNN)-based FER methods, which extract a hierarchy of nonlinear facial features through multi-layers of convolution and pooling, can achieve higher rates of accuracy on several facial expression benchmarks. Deep neural networks models such as Deep Belief Network (DBN) and Deep Boltzmann Machine (DBM) have also been applied to FER with success. Deep learning features have proved to be efficient in extracting crucial patterns from images and have better discriminative power to classify into respective emotion classes as compared to handcrafted features. Table 2.3 highlights a few notable works performed using deep learning techniques.

In a constraint environment, Tong et al. (2016) has extracted Scale-Invariant Feature Transform (SIFT) and has fed these features to Deep Neural Network (DNN) to learn discriminative patterns. A combination of CNN and image processing techniques have been utilized by Lopes et al. (2017) to extract expression specific features that have proved to be efficient for FER. Whereas, Barros et al. (2017) has used convolution units of CNN to identify the location of expression in a cluttered scene rather than using it for classification purposes. Likewise, Kim et al. (2017) has utilized CNN to extract spatial features and trained LSTM, thereby generating discriminative spatio-temporal representation to improve FER with varying expression intensities. Similarly, Pan et al. (2019) has aggregated spatial and temporal features using the aggregation layer, thus filling the gap between visual features and emotions. Zhang et al. (2017) has utilized Part-based Hierarchical Bidirectional Recurrent Neural Network (PHRNN) and Multi-Signal CNN (MSCNN) to extract dynamic and morphological variations of facial expressions in video sequences.

To efficiently recognize expressions from facial movements and body gestures, Sun et al. (2018) has utilized CNN, Bilateral Long Short-Term Memory Recurrent Neural

## 2.2. Deep Learning (DL) Approaches used in Constrained and Unconstrained Environments

---

Networks (BLSTM-RNN) and Principal Component Analysis (PCA). Whereas, [Dandan et al. \(2020\)](#) proposed a BiLSTM architecture and fused both spatial and temporal dynamics jointly for FER. Also, [Sun et al. \(2019\)](#) extracted spatial features from the gray-level image and optical flow features extracted from X and Y components of the emotional and neutral face image. The spatial-temporal features extracted from three-channel elements are fused using Multi-channel Deep Spatial-Temporal feature Fusion neural Network (MDSTFN) for FER. [Xie and Hu \(2018\)](#) proposed Deep Comprehensive Multi-patches Aggregation CNN (DCMA-CNN) with Expressional Transformation-Invariant (ETI) pooling to distinguish expression sensitive elements and reduce the negative impact. [Ma et al. \(2019\)](#) has conducted cross-model noise modeling, to eliminate data pollution in audio data and data redundancy in visual data using 2D CNN and 3D CNN, respectively. The Fusion, of uniform LBP and geometric features using autoencoders, has helped [Majumder et al. \(2016\)](#) in representing the non-linear data in lower dimensions. [Tang et al. \(2018\)](#) has proved that automatically extracted features are superior when compared to handcrafted features. On the other hand, [Zhi et al. \(2021\)](#) used an evolutionary DL approach to evaluate AUs, and it proved to be efficient compared to other AU detection algorithms. [Pham et al. \(2019\)](#) utilized Multi-Layer Perceptron (MLP) as a classifier to determine whether the current classification results are reliable or not. In case of unreliability, the facial image is used to search for similar images. Images with identical facial expressions are processed using AUs from a vast set of unlabeled face datasets.

In an unconstrained environment, few DL works have proved to be efficient in solving variations of in-the-wild databases, cross-cultural problems, vast computational complexity, overfitting, and small data sample issues. [Georgescu et al. \(2019\)](#) has used architectures of CNN and Bag-of-Visual-Words (BOVW) handcrafted features to recognize facial expressions and has employed a local and global learning approach using SVM. Amongst them, local learning SVM proved to be efficient for predicting the class label. For recognizing emotions from video sequences [Kaya et al. \(2017\)](#) has proposed a multimodal approach. [Yan et al. \(2018\)](#) has utilized cascaded

CNN and Bidirectional Recurrent Neural Networks (BRNN) to extract discriminative texture features, model temporal relationships between frames, utilized landmark actions using CNN and SVM, and acoustic features using CNN. Further, feature level fusion and decision level fusion strategy solved the issue of recognizing emotions in in-the-wild databases. Whereas, [Ruan et al. \(2021\)](#) proposed Feature Decomposition and Reconstruction Learning (FDRL) method to model expression similarities, characterize the expression-specific variations and reconstruct expression features. [Cai et al. \(2021\)](#) explicitly reduced inter-subject variations created by identity-related face attributes by proposing Identity-Free conditional Generative Adversarial Network (IF-GAN).

[Xie et al. \(2019\)](#) has proposed a Deep Attentive Multi-path Convolutional Neural Network (DAM-CNN) to locate expression-sensitive regions and has generated high-level representations that are robust against variations like gender, races, etc. Likewise, [Li et al. \(2018\)](#), has helped to overcome the problem of facial occlusions and has improved recognition rate on both occluded and non-occluded faces using CNN with Attention (ACNN) mechanism. [Li and Deng \(2018b\)](#) has utilized Deep Locality-Preserving CNN (DLP-CNN) along with Locality Preserving (LP) loss to form a compact intra-class local clusters for faces belonging to the same emotion classes. This approach has been powerful in handling cross-cultural problems. [Sun and Xia \(2020\)](#) has proposed AlexNet, GoogleNet architectures, to improve the accuracy of CNN architecture and has introduced a new augmentation strategy, “artificial face,” to overcome the overfitting problem caused by CNN architecture. Also, [Shao and Qian \(2019\)](#) has proposed three novel CNN architectures to overcome overfitting, high computational complexity, and shortage of training samples. Whereas [Agrawal and Mittal \(2020\)](#) utilized two novel CNN models and varied the kernel size and number of filters, and showed the variation on the metric accuracy. [Khorrami et al. \(2015\)](#) demonstrated that CNNs trained for emotion recognition are able to predict high-level features that firmly match facial AUs both qualitatively and quantitatively. Whereas, [Liu et al. \(2018\)](#) proposed a novel conditional convolutional neural network enhanced random forest (CoNERF) for recognizing FER in an unconstrained

environment, and the proposed model proved efficient in multi-view FER.

[Hung et al. \(2019\)](#) utilized a learning emotion database collected by students of a National University, Taiwan. The authors performed two phases of transfer learning using Dense\_FaceLiveNet to solve the problem of small data for classifying learning-centered emotions. To overcome the issue of vast parameters used for model training, [Riaz et al. \(2020\)](#) has proposed a shallow net known as eXnet. [Liu et al. \(2021\)](#) proposed a dynamically multi-channel metric network (DML-Net) for handling pose-aware and identity-invariant FER, thus overcoming overfitting and vanishing gradient issues and improving the overall performance.

Table 2.3: Summary of literature based on Deep Learning Techniques

RC	Reference	Dataset	Type of Data	Approaches	Remarks
Lab	Tong et al. (2016)	BU-3DFE, Multi-PIE	S	Scale-Invariant Feature Transform (SIFT), Deep Neural Network (DNN)	<ul style="list-style-type: none"> <li>Deals with multi-view FER. The use of landmark points alleviates the misalignment problem.</li> <li>Overcomes the problem of overfitting and reduces the model complexity.</li> <li>Accuracy of proposed work at a 90 deg pose angle is less when compared to Jung et al. (2015) on the Multi-PIE database.</li> <li>Accuracy is low with squint expression class on Multi-PIE and fear expression class on BU-3DFE datasets.</li> </ul>
	Majumder et al. (2016)	MMI, CK+	D	Facial keypoints (Geometric features), Uniform-LBP, autoencoders, Kohonen Self-Organizing Map (SOM), SVM	<ul style="list-style-type: none"> <li>The performance is computationally efficient and accurate. The fusion of geometric and appearance features using autoencoders has provided the best representation of facial attributes to recognize facial expression.</li> <li>The soft thresholding technique used at the SOM classifier's output nodes reduces the problem of misleading class prediction.</li> <li>LBP feature extraction applied to four facial key regions reduces the redundant information.</li> <li>Focus is on high-level semantic concepts of expression, and the proposed method has ignored fine-grained information at local facial regions.</li> <li>The performance is not been investigated in the real-time environment settings. It requires the first frame of sequence to be of neutral expression, which is not always possible to get in real-life databases.</li> </ul>
	Lopes et al. (2017)	CK+, JAFFE, BU-3DFE	S	CNN	<ul style="list-style-type: none"> <li>Alleviates the data shortage problem.</li> <li>The model has utilized less time for training the system.</li> <li>The proposed work is suitable to operate in real-time environment settings.</li> <li>The accuracy is poor for sad expression class as when compared to other emotion classes.</li> <li>The robustness of the model with various head poses has not been verified.</li> </ul>

\* RC–Recording Conditions, IW–In-the-Wild, S–Static, D–Dynamic

RC	Reference	Dataset	Type of Data	Approaches	Remarks
Lab	Zhang et al. (2017)	CK+, Oulu-CASIA, MMI	D	Part-based Hierarchical bidirectional Recurrent Neural Network (PHRNN) & Multi-signal Convolutional Neural Network (MSCNN)	<ul style="list-style-type: none"> <li>• The proposed method increases variations across distinct expressions and reduces the differentiation of within-class expressions.</li> <li>• It lacks in capturing information of expressions when there is a small motion around the critical areas of the face.</li> </ul>
	Kim et al. (2017)	MMI, CASME II	D	CNN and Long Short Term Memory (LSTM)	<ul style="list-style-type: none"> <li>• The proposed approach overcomes the problem of variations in expression intensity and duration of expression.</li> <li>• Utilization of expression-state information (onset, apex, offset) improves FER performance and the recognition rate of microexpressions.</li> <li>• Layers of LSTM need fine-tuning using real-world datasets to improve the recognition rate further.</li> </ul>
	Barros et al. (2017)	FABO	D	Attention Cross-channel Convolution Neural Networks (CCCNN)	<ul style="list-style-type: none"> <li>• The proposed approach with shunting neurons, filters the noise, and tend to learn the most relevant features.</li> <li>• The model accuracy drops down when more than two faces are present in a sequence of images, and fed as input to recognize the location of expression.</li> <li>• Accuracy is low for expressions like happiness, fear, and boredom.</li> </ul>
	Xie and Hu (2018)	CK+, JAFFE	S & D	Deep Comprehensive Multi-patches Aggregation-Convolutional Neural Networks (DCMA-CNN)	<ul style="list-style-type: none"> <li>• The focus provided on high-level semantic information from a holistic region and fine-grained information from the local areas has helped in getting an improved performance.</li> <li>• Expressional Transformation Invariant (ETI) pooling handles variations like noises, illumination, image rotations, and reduced negative impact.</li> <li>• ETI pooling has enhanced the discriminative ability of a model and has helped distinguish sensitive expression elements by fusing different features.</li> <li>• Misclassification is highest for sadness emotion class of CK+ and fear emotion class of JAFFE datasets.</li> </ul>

\* RC–Recording Conditions, IW–In-the-Wild, S–Static, D–Dynamic

RC	Reference	Dataset	Type of Data	Approaches	Remarks
Lab	Sun et al. (2018)	FABO	D	CNN, Bilateral Long Short-Term Memory Recurrent Neural Networks (BLSTM-RNN), PCA, SVM	<ul style="list-style-type: none"> <li>The onset-apex-offset video skeleton strategy for frame sequence extraction gives the best recognition rate.</li> <li>The model is not much robust against noisy and stable illumination conditions.</li> <li>The accuracy obtained from video-words (images) are less, when compared to video-skeleton (key clips).</li> </ul>
	Tang et al. (2018)	CK+, Oulu-CASIA, MMI	D	Geometric features, Artificial Neural Networks (ANN), CNN	<ul style="list-style-type: none"> <li>The Learning Propagation (LP) method used to fuse geometric and automatically extracted features gives the best results.</li> <li>Tanh activation function applied at the last second layer of ANN architecture helps overcome the gradient explosion and gradient disappearance problem of ANN.</li> <li>Low accuracy has been attained with handcrafted features.</li> </ul>
	Pham et al. (2019)	FER 2013	S	Multi-Layer Perceptron (MLP), GoogLeNet, Densely Connected Convolutional Networks (DenseNet), Visual Geometry Group-Face (VGG-Face)	<ul style="list-style-type: none"> <li>The focus of this work is to check whether the FER result is reliable or not. The authors evaluate whether further information is needed to make a reclassification to improve FER performance based on the reliability measure.</li> <li>Considering uncertainty with MLP and emotion-preserving image retrieval with AUs increased the network's performance.</li> <li>The proposed technique can be combined with any Deep Learning architecture to boost the system's performance even more.</li> </ul>
	Pan et al. (2019)	RML, eINTERFACE05	D	CNN (VGG-19) & LSTM	<ul style="list-style-type: none"> <li>The proposed framework can extract comprehensive features efficiently using aggregation of spatial and temporal approaches, thus fix the gap between visual features and emotions.</li> <li>Expensive, as it uses a massive number of network parameters and consumes vast computation time.</li> <li>Classification of fear emotion class is poor on both datasets.</li> </ul>

\* RC–Recording Conditions, IW–In-the-Wild, S–Static, D–Dynamic



RC	Reference	Dataset	Type of Data	Approaches	Remarks
Lab	Ma et al. (2019)	RML, eINTERFACE05, BAUM-1s	D	Audio network: 2-Dimensional Convolutional Neural Networks (2D CNN); Video network: 3-Dimensional Convolutional Neural Networks (3D CNN), DBN, SVM	<ul style="list-style-type: none"> <li>Solves the issue of data redundancy and data pollution (denoising) by considering cross-modal feature fusion.</li> <li>The data preparation process takes a lot of time, making the real-time performance of the system more miserable.</li> </ul>
	Sun et al. (2019)	CK+, MMI, RaFD	S	Multi-channel Deep Spatial-Temporal feature Fusion neural Network (MDSTFN)	<ul style="list-style-type: none"> <li>Optical flow information proved to be an effective supplement to spatial features in improving FER's performance from static images.</li> <li>The difference between one neutral-face image and the emotional-face image is employed to compute optical flow extraction instead of evaluating the consecutive sequence of images.</li> </ul>
	Dandan et al. (2020)	CK+, Oulu-CASIA, MMI	D	Deep Spatial Network (DSN) + Deep Temporal Network (DTN) + Bidirectional Long Short-Term Memory (BiLSTM) (Inception-w is utilized as a basic network)	<ul style="list-style-type: none"> <li>Discriminative spatial features are crucial for FER.</li> <li>Average accuracy on Oulu-CASIA and MMI is decreased when convolutional layers were exceeded by three.</li> </ul>
	Zhi et al. (2021)	DISFA, BP4D-Spontaneous	D	3D convolutional neural network (3DLeNet), Boundary Equilibrium Generative Adversarial Networks (BEGAN), Genetic Algorithm	<ul style="list-style-type: none"> <li>Among the numerous AU detection algorithms, EvoNet produced the best results. With BEGAN approach, training was easier, and it converges stably.</li> <li>The model is robust and provides better generalization.</li> <li>Training of Deep Neural networks is expensive; hence, identifying proper generation parameters and initial population numbers to avoid hardware limitation is necessary.</li> </ul>

\* RC–Recording Conditions, IW–In-the-Wild, S–Static, D–Dynamic

RC	Reference	Dataset	Type of Data	Approaches	Remarks
Lab & IW	Khorrani et al. (2015)	CK+, TFD	S	Zero-bias CNN with Data Augmentation and Dropout (AD)	<ul style="list-style-type: none"> <li>The network was trained quickly, and the number of parameters used to learn was lowered simultaneously, bypassing the bias.</li> <li>When visualizing discriminative spatial patterns, the authors found that most of the filters are excited by the face regions that correspond to Facial AUS.</li> </ul>
	Kaya et al. (2017)	EmotiW 2015 & 2016 challenge datasets, ChaLearn-LAP first impressions challenge dataset, RECOLA, CK+, MMI	D	SIFT, HOG, Local Phase Quantization (LPQ), LBP, Local Gabor Binary Pattern-Three Orthogonal Planes (LGBP-TOP), deep CNN, Extreme Learning Machine (ELM), Partial Least squares (PLS) regression.	<ul style="list-style-type: none"> <li>A combination of strategies like weight decay and dropouts with regularization gives the best results.</li> <li>Usage of pre-trained CNN models complement systems with various modalities and helps in efficient feature extraction.</li> <li>The performance of multimodal systems is poor compared to unimodal for arousal predictions on the RECOLA dataset.</li> <li>Emotion classes like disgust and fear give low results on the EmotiW challenge dataset.</li> </ul>
	Li and Deng (2018b)	RAF-DB, CK+, MMI, SFEW 2.0	S	Deep Locality-Preserving Convolutional Neural Network (DLP-CNN)	<ul style="list-style-type: none"> <li>DLP-CNN with Locality Preserving (LP) loss obtains more discriminative features, which improves the recognition and enhances the system's classification performance.</li> <li>LP loss forms a good and compact intra-class local cluster for each category. The execution time of the proposed method using LP takes more time when compared to the center loss.</li> </ul>
	Liu et al. (2018)	CK+, JAFFE, multi-view BU-3DEF, LFW	S	Conditional Convolutional Neural Network Enhanced Random Forest (CoNERF)	<ul style="list-style-type: none"> <li>The influence of distortion induced in an unconstrained context is avoided by considering salient features retrieved from saliency-guided face patches.</li> <li>The deep salient features contribute to a more accurate description of multi-view facial expression images.</li> <li>The proposed learning strategy utilizes a global deep salient representation. Unlike other deep neural networks that require large amounts of training data, conditional CoNERF works effectively even with a small amount of training data.</li> </ul>

\* RC–Recording Conditions, IW–In-the-Wild, S–Static, D–Dynamic

RC	Reference	Dataset	Type of Data	Approaches	Remarks
Lab & IW	Li et al. (2018)	FED-RO, RAF-DB, AffectNet, CK+, MMI, Oulu-CASIA, SFEW	S	CNN with attention mechanism (ACNN)	<ul style="list-style-type: none"> <li>The issues like partial occlusions are solved using an attention mechanism.</li> <li>The performance has a negative impact due to the misalignment of facial landmark points.</li> <li>Global-local-based ACNN (gACNN) suffers from extremely severe facial occlusions and real occlusions.</li> </ul>
	Xie et al. (2019)	CK+, JAFFE, TFEID, SFEW, FER2013, BAUM-2i	S	VGG-Face network, Salient Expressional Region Descriptor (SERD) and Multi-Path Variation-Suppressing Network (MPVS-Net)	<ul style="list-style-type: none"> <li>Deep Attentive Multipath Convolutional Neural Network (DAM-CNN) jointly uses SERD and MPVS-Net; it can learn discriminative features and perform well in constraint and unconstrained environment.</li> <li>The model overcomes severe overfitting issues. The dropout layer added to the model helps in partly improving the generalization ability of the model.</li> <li>The SERD approach may fail to focus on salient facial regions, due to the vast variations in unconstrained datasets.</li> <li>MPVS-Net (autoencoders and decoders) uses a massive amount of parameters and requires further exploration.</li> </ul>
	Shao and Qian (2019)	CK+, BU-3DFE, FER2013	S	Multi-Task Convolutional Neural Network (MTCNN), LBP, CNN	<ul style="list-style-type: none"> <li>The model overcomes the issues of deep CNN like overfitting, high computational complexity, and insufficient data sample.</li> <li>A shallow network with few parameters has proved to be efficient on all three datasets.</li> <li>The performance of FER2013 is lower with the combination of LBP and deep CNN architecture.</li> <li>Expression classes like sadness and surprise gave poor accuracy on the FER2013 dataset.</li> </ul>

\* RC–Recording Conditions, IW–In-the-Wild, S–Static, D–Dynamic

RC	Reference	Dataset	Type of Data	Approaches	Remarks
Lab & IW	Georgescu et al. (2019)	FER2013, FERPlus, AffectNet	S	SIFT descriptors and k-means clustering (computed by Bag of Visual Words (BoVW) model); CNN architectures (VGG-face, VGG-f and VGG-13), k-Nearest Neighbor (k-NN), SVM	<ul style="list-style-type: none"> <li>Overcomes the overfitting problem of CNN using Dense-Sparse-Dense (DSD) approach during training.</li> <li>Local learning with the SVM approach proved to be efficient when compared to global learning.</li> <li>The proposed approach cannot distinguish between voluntary and involuntary facial expressions.</li> </ul>
	Hung et al. (2019)	JAFFE, CK+, FER2013, Learning emotion database	S & D	Dense FaceLiveNet	<ul style="list-style-type: none"> <li>Transfer learning approach solved the problem of a small number of data samples present in a learning-centered emotion dataset.</li> <li>The model does not overcome exceptional real-time situations like occlusions, illumination variations, which can occur in a real-time classroom environment.</li> </ul>
	Sun and Xia (2020)	CK+, JAFFE, FER2013, self-collected database	S	CNN (AlexNet, GoogleNet)	<ul style="list-style-type: none"> <li>The proposed approach increases speed, accuracy, and reduces computational complexity.</li> <li>Overcomes data overfitting problems using an artificial face augmentation strategy.</li> <li>The model is not robust towards the in-the-wild database, and the recognition rate is lesser compared to the controlled environment.</li> <li>The performance of AlexNet and GoogleNet architectures on anger expression class is lower.</li> <li>The choice of kernel size for choosing the mask needs to be carefully selected. Otherwise, it may lead to the wrong selection of Region of Interest (ROI) and further decrease the system's overall performance.</li> </ul>

\* RC–Recording Conditions, IW–In-the-Wild, S–Static, D–Dynamic

RC	Reference	Dataset	Type of Data	Approaches	Remarks
Lab & IW	Riaz et al. (2020)	CK+, FER2013, RAF-DB	S	CNN (Expression Net)	<ul style="list-style-type: none"> <li>• A lightweight network with a small number of parameters reduces overhead.</li> <li>• Employing <math>1 \times 1</math> convolutions before <math>3 \times 3</math> convolutions decreases the model's size while maintaining the model's accuracy.</li> <li>• The model overcomes the overfitting issue.</li> <li>• The model's efficiency needs to be improved for recognition of facial expression in an unconstrained environment.</li> </ul>
	Ruan et al. (2021)	CK+, MMI, Oulu-CASIA, RAF-DB, SFEW	S & D	Feature Decomposition and Reconstruction Learning (FDRL) (Backbone network utilized is Resnet18)	<ul style="list-style-type: none"> <li>• The proposed FDRL model accurately identifies the expression similarities, expression-specific variations and extraction of fine-grained expression features.</li> <li>• There exists a redundancy and noise in latent features, which requires further exploration.</li> </ul>
	Cai et al. (2021)	BU-3DFE, CK+, MMI, RAF-DB	S & D	Identity-Free conditional Generative Adversarial Network (IF-GAN) consisting of U-Net and Patch Generative Adversarial Network (PatchGAN); ResNet-101 as expression classifier.	<ul style="list-style-type: none"> <li>• The proposed IF-GAN alleviates identity-related information and produces a facial image for FER that is identity-free.</li> <li>• IF-GAN can overcome pose, occlusions and illumination variations.</li> </ul>
	Liu et al. (2021)	KDEF, BU-3DFE, Multi-PIE, SFEW 2.0	S	Dynamically Multi-channel Metric-Network (DML-Net)	<ul style="list-style-type: none"> <li>• The DML-Net reduces deep multiple metric loss, FER loss, and pose-estimation loss by employing dynamically learned loss weights, reducing overfitting, and enhancing recognition significantly.</li> <li>• To achieve robust FER performance, the DML-Net reduces the effects of pose and identity.</li> </ul>

\* RC–Recording Conditions, IW–In-the-Wild, S–Static, D–Dynamic

RC	Reference	Dataset	Type of Data	Approaches	Findings
Lab & IW	Liu et al. (2020)	RAF-DB 2.0, FER2013, CK+	S	Point Adversarial Self Mining (PASM) (backbone network used: ResNet-34 and VGG-16).	<ul style="list-style-type: none"> <li>The PASM model is proposed for data augmentation and modeling the data distribution in the wild. It helps locate the sample's most informative region via the point adversarial attack policy.</li> <li>By taking into account how different regions contribute to classification in each sample according to the model's specifications, the PASM model is able to self-mine knowledge from provided data.</li> <li>Searching for a sensitive position in each image is time-consuming. Choice of proper iteration number is expected for PASM to work efficiently.</li> </ul>
	Koujan et al. (2020)	FaceVid, Radboud, KDEF, RAF-DB, CFEE, CK+	S & D	Deep-Exp3D, SVM.	<ul style="list-style-type: none"> <li>The network is robust against viewing angle and illumination variations, occlusions, and regresses the expression independently irrespective of the persons' identity.</li> </ul>
	Liang et al. (2020)	FG-Emotions, CK+, MMI, Oulu-CASIA, AFEW, SFEW, RAF-DB, AffectNet, FER-2013	S & D	Multi-Scale Action Unit (AU)-based Network (MSAU-Net) for recognition of images and TMSAU-Net with attention mechanism and temporal stream to jointly learn spatial and temporal features.	<ul style="list-style-type: none"> <li>The mouth region needs to be selected when emotion is optimistic, and areas like eyes and eyebrow regions need to be chosen when emotion is passive.</li> <li>The facial muscle movements of positive expressions are large than negative expressions, and the accuracy is also higher.</li> </ul>
	Saurav et al. (2022)	FER2013, FERPlus, RAF-DB, CK+	S	Dual Integrated Convolution Neural Network (DICNN)	<ul style="list-style-type: none"> <li>DICNN models overcome the issue of computationally intensive and extensive memory storage by using two custom lightweight CNNs.</li> <li>The proposed model utilizes 1.08M model parameters and 5.40MB storage memory and provides the best tradeoff between recognition accuracy and computational efficiency.</li> <li>The proposed system is ideal for real-world applications since it runs in real-time on a resource-constrained embedded platform.</li> </ul>

\* RC–Recording Conditions, IW–In-the-Wild, S–Static, D–Dynamic

RC	Reference	Dataset	Type of Data	Approaches	Findings
IW	Yan et al. (2018)	AFEW 6.0, CHEAVD	D	CNN, Bidirectional Recurrent Neural Networks (BRNN), SVM	<ul style="list-style-type: none"> <li>The emotions from in-the-wild datasets are recognized efficiently using three cues like facial texture, facial landmark action, and audio signal.</li> <li>The recognition from audio signals is deprived, as it contains noisy data.</li> <li>Recognition of anxious and worried emotion classes is lower, and the highest misclassification is observed in disgust emotion class due to imbalanced and fewer data samples present in these emotion categories.</li> </ul>
	Nguyen et al.	FER2013, AFEW 7.0	S & D	Ensemble of Multi-Level Convolutional Neural Network (MLCNN) and temporal model with an ensemble of MLCNN and 3DCNN.	<ul style="list-style-type: none"> <li>The addition of mid-level and high-level features from few blocks played a vital role in the classification task.</li> <li>The filter size 3x3 proved to be performing well on most of the image classification problems.</li> </ul>
	Li et al. (2019)	RAF-DB, AffectNet	S	Resnet-18 with separate loss and softmax loss.	<ul style="list-style-type: none"> <li>The softmax loss layer alone is not sufficient enough to discriminate facial expression recognition on an in-the-wild dataset.</li> <li>A separate loss function is required along with the softmax layer to recognize basic and compound expressions efficiently.</li> </ul>
	Xiaohua et al. (2019)	AffectNet	S	Residual attention block, Bi-directional Recurrent Neural Network (Bi-RNN) with self-attention.	<ul style="list-style-type: none"> <li>The two-level attention block yielded the best results. But, the performance of valence was poor as compared to arousal.</li> <li>Tukey's biweight loss function was utilized to reduce the impact of erroneous samples.</li> </ul>
	Agrawal and Mittal (2020)	FER-2013	S	CNN	<ul style="list-style-type: none"> <li>Metric accuracy is unstable for very low and very high kernel sizes.</li> </ul>

\* RC–Recording Conditions, IW–In-the-Wild, S–Static, D–Dynamic

RC	Reference	Dataset	Type of Data	Approaches	Findings
IW	Reddy et al. (2020)	AffectNet	S	Faster Region based Convolutional Neural Network (RCNN) was used to extract face regions; Facial landmark points and XceptionNet features were extracted; SVM with Radial Basis Function (RBF) was used for classification.	<ul style="list-style-type: none"> <li>Deep learning methods fail in an unconstrained situation, as image background dominates over the facial features. Thus, a combination of deep learning and machine learning features is essential in solving such issues.</li> </ul>
	Yan et al. (2020)	RAF-AU	S	AU detection with CNN (AU-CNN).	<ul style="list-style-type: none"> <li>It is challenging to categorize expressions in an in-the-wild dataset based on the AU patterns.</li> <li>The annotation of facial expression should include both subjective (judgement-based) and objective (sign-based) elements.</li> <li>Facial action units differ from one side of the face to another, they are not symmetrical, and they vary in the intensity value.</li> </ul>
	Wang et al. (2020a)	RAF-DB, FERPlus, AffectNet, WebEmotion	S & D	Self-Cure Network (SCN) utilized ResNet 18 as a backbone network.	<ul style="list-style-type: none"> <li>The proposed SCN reduces the uncertainty caused by ambiguous facial expressions, low quality images and subjectiveness of annotators for large scale FER.</li> </ul>
	Pengcheng et al. (2020)	EmotioNet 2020	S	Multi-view co-regularization framework with checkpoints and threshold for AU recognition.	<ul style="list-style-type: none"> <li>By choosing the optimal checkpoint for each AU, recognition can be improved as AUs converge at varying speeds.</li> <li>The multi-view and the co-regularization loss benefit the supervised training; also, results are better than semi-supervised training.</li> </ul>
	Wang et al. (2020b)	FERPlus, RAF-DB, AffectNet, and SFEW	S	Region Attention Network (RAN) to obtain important facial regions, Region Biased Loss (RB-Loss) was used to assign a high weight to the most important region.	<ul style="list-style-type: none"> <li>RAN can effectively capture the action units related to expressions surprise, happiness, and sadness (Cheek raiser, lip corner puller, lip corner depressor).</li> <li>RAN improves the performance of recognition in varying conditions like occlusions and pose variations.</li> </ul>

\* RC–Recording Conditions, IW–In-the-Wild, S–Static, D–Dynamic



RC	Reference	Dataset	Type of Data	Approaches	Findings
IW	Zhao et al. (2020)	FER2013, FERPlus and FERFIN	S	Lightweight Emotion Recognition (LER) utilizing DenseNet architecture.	<ul style="list-style-type: none"> <li>The proposed LER model addresses the latency in natural conditions and eliminates redundant parameters.</li> <li>The model showed poor recognition on few primary expression classes that need further exploration.</li> </ul>
	Zhu et al. (2020b)	RAF-DB, AffectNet	S	ResNet-18 is used as backbone for Deep Locality Preserving-Convolutional Neural Networks (DLP-CNN) network.	<ul style="list-style-type: none"> <li>The proposed center-expression-distilled loss improved the discriminative quality of deeply learned features and avoided catastrophic forgetting.</li> <li>The new dimension added at the fully connected (FC) layer assigns higher prediction scores to the new expression classes than the old one using incremental learning.</li> <li>Solves the problem of consuming large computation resources.</li> </ul>
	Vo et al. (2020)	RAF-DB, AffectNet, FERPlus	S	Pyramid With Super Resolution (PSR), Backbone Network: VGG-16	<ul style="list-style-type: none"> <li>PSR deals with varying image size problems.</li> <li>The Super Resolution (SR) methods applied to upscale the low-resolution input images improved the network performance.</li> <li>By utilizing prior knowledge of the confusion about each expression, the Prior Distribution Label Smoothing (PDLs) loss function enhanced the FER problem.</li> </ul>
	Farzaneh and Qi (2021)	AffectNet, RAF-DB	S	Resnet-18, Deep Attentive Center Loss (DAKL)	<ul style="list-style-type: none"> <li>To improve feature discrimination, DAkl can be used in conjunction with other classification tasks.</li> <li>More research into reducing primary emotion misclassification in in-the-wild database is required.</li> </ul>
	Sepas-Moghaddam et al. (2021)	LFFW, LFFC	S	Resnet-50 + capsule network; VGG-16 + Capsule network	<ul style="list-style-type: none"> <li>The capsule network adds network value in learning a model that completely utilizes the angular features available in Light Field images.</li> <li>The proposed CapsField requires more training time to extract more discriminative features.</li> </ul>

\* RC–Recording Conditions, IW–In-the-Wild, S–Static, D–Dynamic

RC	Reference	Dataset	Type of Data	Approaches	Findings
IW	Chen et al. (2021a)	300-W, AFLW, AffectNet, RAF-DB	S	VGG-16 and ResNet 50 as a backbone network.	<ul style="list-style-type: none"> <li>Residual multi-task learning framework is proposed to carry out landmark localization and expression recognition tasks.</li> <li>Association learning method is further proposed to enhance the two tasks.</li> <li>The proposed models' speed is not fast enough to make it deploy for real-world applications.</li> </ul>
	Zhu et al. (2022)	RAF-DB, SFEW, FER2013	S	Convolutional Relation Network (CRN)	<ul style="list-style-type: none"> <li>Few-shot learning is incorporated into the proposed method for transferring discriminative information to determine new emotion classes.</li> <li>Overcame the issue of class imbalance, which is a major challenge in the In-the-wild field.</li> </ul>
	Nan et al. (2022)	RAF-DB, SFEW 2.0	S	Feature level super-resolution method for robust facial expression recognition (FSR-FER)	<ul style="list-style-type: none"> <li>The proposed FSR-FER lowers the risk of privacy leaking without recovering high-resolution facial images.</li> <li>The performance is better on low-resolution images with more feature loss.</li> <li>The proposed approach overcomes the FER problem of multi-facial images in crowd scenarios.</li> <li>The introduction of classification-aware loss reweighting into FSR-FER achieves faster training convergence and better performance.</li> </ul>

\* RC–Recording Conditions, IW–In-the-Wild, S–Static, D–Dynamic

### 2.3 TRADITIONAL MACHINE LEARNING VERSUS DEEP LEARNING TECHNIQUES

The transition from traditional ML to DL algorithms has been ascertained quite effectively for detecting FER on both static and dynamic data. ML models have proven to be efficient when there are fewer data and are trained on fewer parameters; thus, it is unsuitable for large datasets. ML models take handcrafted features as input which is difficult to generalize (Hung et al. 2019). Whereas DL contains multiple intermediate layers between the input and classification layers, it can automatically learn higher-level semantic characteristics from vast training data (Liu et al. 2018; Mollahosseini et al. 2016; Sun et al. 2019). Using many layers of convolution and pooling layers, the CNN network can extract non-linear face features (Ge et al. 2022; Minaee et al. 2019), resulting in more excellent identification rates on the FER system. Although DL models are computationally expensive, they aid in acquiring complex features, and the CNN model's generalization ability is better than traditional ML models (Hung et al. 2019).

### 2.4 SUMMARY OF FER TECHNIQUES BASED ON THE DATA

The FER systems can be divided into two categories, the work considering the static images and those that work with dynamic image sequences (Kim et al. 2017). In static based approach, the feature vector contains the information about the current image and overlooks the temporal information (Lopes et al. 2017). On the other hand, the dynamic sequence-based method utilizes the temporal information between one or more frames to recognize the facial expressions. The FER automated systems take static or dynamic images as input and classify the expressions for controlled or uncontrolled scenario data. The dynamic sequence data can extract spatial and temporal characteristics and achieve better performance, but it leads to computational complexity and introduces noise and disruption (Sun et al. 2019). Table 2.4 summarizes the FER techniques based on the type of the data.

Table 2.4: Summary of FER Techniques based on the Type of Data

Type	Type of Data	FER Techniques	Performance Metric
Traditional	S	LBP, NCM, SVM (Ghimire et al. 2017)	ACC: 97.25% (CK+)
		LBP-TOP, Bi-WOOF, SVM (Liong et al. 2018)	ACC: 58.85% (CASME II), 62.20% (SMIC-HS)
		Gabor filter, PCA, Genetic Algorithm (Boughida et al. 2022)	ACC: 96.30% (JAFFE), 94.20% (CK), and 94.26% (CK+)
		LBP, LBP-TOP, Group LASSO, SVD (Jia et al. 2018)	ACC: 65.5% (CASMEII)
	D	Uniform LBP, SVM (Zhong et al. 2014)	ACC: 91.53% (CK+), 77.39% (MMI), 80% (GEMEP-FERA)
		LBP, SVM (Cruz et al. 2014)	ACC: 71.8% (MMI), 56% (AVEC 2011), 76.1% (CK+)
		HOG-TOP, geometric warp feature, SVM (Chen et al. 2016)	ACC: 95.7% (CK+), 45.2% (AFEW 4.0)
		ELBPTOP (Guo et al. 2019)	ACC: 73.94% (CASME II), 69.06% (SMIC) [with original classes]; 63.44% (SAMM), 79.55% (CASME II) [reorganized classes]
	S & D	LBP, ORB, SVM (Niu et al. 2021)	ACC: 88.5% (JAFFE), 93.2% (CK+), 79.8% (MMI)
	Deep Learning	S	SIFT, DNN (Tong et al. 2016)
CNN (Lopes et al. 2017)			ACC: 96.76%
MLP, GoogLeNet, DenseNet, VGG-Face (Pham et al. 2019)			ACC: 69.18% (FER 2013)
MDSTFN (Sun et al. 2019)			ACC: 98.38% (CK+), 98.75% (RaFD), 99.59% (MMI)
Zero-bias CNN with Data Augmentation and Dropout (AD) (Khorrami et al. 2015)			ACC: 89.8% (TFD), 96.4% (CK+)
DLP-CNN (Li and Deng 2018b)			ACC: 95.78% (CK+), 51.05% (SFEW 2.0), 78.46% (MMI), 74.20% (RAF-DB (basic)), 44.55% (RAF-DB (compound))
Conditional CoNERF (Liu et al. 2018)			ACC: 94.09% (Multi-View BU-3DFE), 99.02% (CK+ and JAFFE), 60.9% (LFW)
ACNN (Li et al. 2018)			ACC: 66.50% (FER-RO)

\* S–Static, D–Dynamic, ACC–Accuracy

## 2.4. Summary of FER techniques based on the Data

Type	Type of Data	FER Techniques	Performance Metric
Deep Learning	S	MTCNN, LBP, CNN (Shao and Qian 2019)	ACC: 95.29% (CK+), 86.50% (BU-3DFE), 71.14% (FER2013)
		VGG-Face network, SERD, MPVS-Net (Xie et al. 2019)	ACC: 95.88% (CK+), 99.32% (JAFFE), 93.36% (TFEID), 61.52% (BAUM-2i), 66.20% (FER2013), 42.30% (SFEW)
		SIFT descriptors, k-means clustering; VGG-face, VGG-f and VGG-13, k-NN, SVM (Georgescu et al. 2019)	ACC: 75.42% (FER2013), 87.76% (FERPlus), 63.31% (AffectNet)
		AlexNet, GoogleNet (Sun and Xia 2020)	ACC: 94.67% (CK+), 53.77% (CK+ → JAFFE), 39.13% (CK+ → FER2013), 36.25% (CK+ → IW)
		CNN (Expression Net) (Riaz et al. 2020)	ACC: 73.54% (FER2013), 96.75% (CK+), 86.37% (RAF-DB)
		DML-Net (Liu et al. 2021)	ACC: 88.2% (KDEF), 83.5% (BU-3DFE), 93.5% (Multi-PIE), 54.39% (SFEW)
		Resnet-18 (Li et al. 2019)	ACC: 86.38% (RAF-DB-basic), 58.84% (RAF-DB-compound), 58.89% (AffectNet)
		Bi-RNN (Xiaohua et al. 2019)	ACC: 48% (AffectNet)
		PASM (Liu et al. 2020)	ACC: 88.68% (RAF-DB) and 73.59% (FER2013)
		CNN (Agrawal and Mittal 2020)	ACC: Model 1- 65.77% and Model 2- 65.23% (FER-2013)
		Faster RCNN; Facial landmark points and XceptionNet features; SVM with RBF (Reddy et al. 2020)	ACC: 59% (AffectNet)
		AU-CNN (Yan et al. 2020)	Area under the ROC Curve (AUC): 88.73; F1-score: 65.95 (RAF-AU)
		Multi-view co-regularization framework (Pengcheng et al. 2020)	ACC: 73.06% (EmotioNet 2020)
		RAN (Wang et al. 2020b)	ACC: 89.16% (FERPlus), 86.9% (RAF-DB), 59.5% (AffectNet), 56.4% (SFEW)
		LER (Zhao et al. 2020)	ACC: 71.73% (FER2013); 85.58% (FERPlus); 85.89% (FERFIN)
DLP-CNN (Zhu et al. 2020b)	ACC: 80.60% (RAF-DB), 82.17% (AffectNet)		
PSR (Vo et al. 2020)	ACC: 88.98% (RAF-DB-Weighted Accuracy (WA)), 80.78% (RAF-DB-Unweighted Accuracy (UA)); 89.75% (FERPlus); 63.77% (AffectNet)		

\* S–Static, D–Dynamic, ACC–Accuracy

## 2. Literature Review

Type	Type of Data	FER Techniques	Performance Metric
Deep Learning	S	DAFL (Farzaneh and Qi 2021)	ACC: 65.20% (AffectNet), 87.78% (RAF-DB)
		Resnet-50 and VGG-16 with capsule network (Sepas-Moghaddam et al. 2021)	ACC: 61.59% (LFFW), 88.25% (LFFC)
		VGG-16, ResNet 50 (Chen et al. 2021a)	Normalized mean errors (NME): 3.49 (300-W), 1.69 (AFLW)
		CRN (Zhu et al. 2022)	ACC: 56.25% (RAF-DB), 67.32% (FER2013), 54.87% (SFEW)
		DICNN (Saurav et al. 2022)	ACC: 72.77% (FER2013), 85.29% (FERPlus), 86.07% (RAF-DB)
		FSR-FER (Nan et al. 2022)	ACC: 76.66% (RAF-DB), 55.14% (SFEW)
	D	Facial keypoints, Uniform-LBP, Autoencoders, SOM, SVM (Majumder et al. 2016)	ACC: 97.55% (MMI), 98.95% (CK+)
		PHRNN, MSCNN (Zhang et al. 2017)	ACC: 98.50% (CK+), 86.25% (Oulu-CASIA), 81.18% (MMI)
		CNN, LSTM (Kim et al. 2017)	ACC: 69.94% (MMI), 58.54% (CASME II)
		Attention CCCNN (Barros et al. 2017)	ACC: 95.13% (FABO)
		CNN, BLSTM-RNN, PCA, SVM (Sun et al. 2018)	ACC: 99.57% (FABO)
		Geometric features, ANN, CNN (Tang et al. 2018)	ACC: 98.73% (CK+), 87.50% (Oulu-CASIA)
		VGG-19, LSTM (Pan et al. 2019)	ACC: 65.72% (RML), 42.98% (eNTERFACE05)
		2D CNN, 3D CNN, DBN, SVM (Ma et al. 2019)	ACC: 82.38% (RML), 85.69% (eNTERFACE05), 59.17% (BAUM-1s)
		DSN+DTN+BiLSTM (Dandan et al. 2020)	ACC: 99.6% (CK+), 91.07% (Oulu-CASIA), 80.71% (MMI)
		3DLeNet, BEGAN, Genetic Algorithm (Zhi et al. 2021)	ACC: 86.3% (BP4D)
		SIFT, HOG, LPQ, LBP, LGBP-TOP, deep CNN, ELM, PLS regression (Kaya et al. 2017)	ACC: 54.55% (EmotiW 2015), 52.11% (EmotiW 2016)
		CNN, BRNN, SVM (Yan et al. 2018)	ACC: 55.14% (CHEAVD), 49.22% (AFEW 6.0)

\* S–Static, D–Dynamic, ACC–Accuracy

2.5. Summary of research works on FER, based on posed, non-posed expressions, facial occlusions and engagement recognition

Type	Type of Data	FER Techniques	Performance Metric
Deep Learning	S & D	DCMA-CNN (Xie and Hu 2018)	ACC: 93.46% (CK+), 94.75% (JAFPE)
		Dense FaceLiveNet (Hung et al. 2019)	ACC: 90.97% (JAFPE), 95.89% (KDEF), 69.99% (FER2013), 79.03% (Learning emotion database), 70.02% (KDEF→FER2013), 91.93% (FER2013→ Learning emotion database)
		FDRL (Ruan et al. 2021)	ACC: 89.47% (RAF-DB), 62.16% (SFEW), 99.54% (CK+), 85.23% (MMI), 88.26% (Oulu-CASIA)
		IF-GAN, PatchGAN, ResNet-101 (Cai et al. 2021)	ACC: 88.33% (RAF-DB), 85.25% (BU-3DFE), 97.52% (CK+), 75.48% (MMI)
		MLCNN, 3DCNN (Nguyen et al.)	ACC: 74.09% (FER2013), 49.3% (AFEW)
		Deep-Exp3D, SVM (Koujan et al. 2020)	ACC: 87.98% (FaceVid)
		SCN (Wang et al. 2020a)	ACC: 88.14% (RAF-DB), 60.23% (AffectNet), 89.35% (FERPlus)
		MSAU-Net, TMSAU-Net (Liang et al. 2020)	ACC: 73.73%, 65.86% (FG-Emotions)

\* S–Static, D–Dynamic, ACC–Accuracy

**2.5 SUMMARY OF RESEARCH WORKS ON FER, BASED ON POSED, NON-POSED EXPRESSIONS, FACIAL OCCLUSIONS AND ENGAGEMENT RECOGNITION**

**2.5.1 Classification of Non-posed Expressions**

Several FER systems have been developed by researchers for the recognition of micro-expressions in the past decades. Liong et al. (2018) inspired to utilize onset and apex frames for recognition of ME rather than using the entire video sequences. The author followed a divide and conquer strategy for spotting the apex frames and extracted Bi-Weighted Oriented Optical Flow (Bi-WOOF) features from the onset and peak frames, which helped efficiently recognize ME's. Also, Kotsia et al. (2006) proposed a method for recognition of FER from video frames by fusing geometrical information and texture information; texture information is obtained using Discriminant Non-negative Matrix Factorization (DNMF). Wang et al. (2014) proposed Tensor Independent Color Space (TICS) and extracted LBP-TOP features to

enhance the performance of MER and utilized SVM for classification. Also, [Thu Nguyen et al. \(2021\)](#) fused the optical flow and dynamic image computation features and fed them into deep learning models like VGG-19, Resnet 50, Inception V3, and EfficientNet B0 to enhance the performance of ME recognition. Whereas [Tang et al. \(2018\)](#) extracted geometric features and utilized Artificial Neural Network (ANN) for classification and also utilized automatic features extracted from CNN. Fusion of geometric and learned features gave the best results.

[Sultan Zia et al. \(2018\)](#) developed a dynamically weighted majority voting (DWMV) strategy. Thus, DWMV helped to recognize spontaneous expressions in varying environment constraints and varying ethnicities and cultures. Multi-Region Ensemble CNN (MRE-CNN) model is proposed by [Fan et al. \(2018\)](#), which is a combination of three different networks for obtaining information from three distinct sub-regions (left eye, nose, and mouth) of the face. The final prediction is based on the weighted sum operation from three prediction scores from each of these networks. This further improves the overall accuracy of FER. To integrate the AU recognition task and exploit AU relational information for recognition of MER, [Xie et al. \(2020\)](#) proposed an AU-assisted Graph Attention Convolutional Network (AU-GACN). The authors have also exploited AU Intensity Controllable Generative Adversarial Nets (AU-ICGAN) to generate a large number of synthetic training samples to overcome the issue of the limited dataset for MER. To integrate the covariance correlation, as well as find discriminative information on limited ME datasets without the use of extra information [Li et al. \(2021\)](#) proposed Spatial Attention Module (SA) and Channel Attention Module (CA).

[Lu et al. \(2015\)](#) encoded texture variations corresponding to muscle movements and proposed a Delaunay-based Temporal Coding Model (DTCM) for recognition of micro-expressions. [Zhang et al. \(2021\)](#) proposed Intradomain Structure Domain Adaptation (IDSDA) to solve Cross-Database Micro-Expression Recognition (CDMER). According to previous research findings, a precise technique is required to record minute face changes and get satisfactory results. Thus, this work proposes a novel MER system using the DT and VD to retrieve the ROI based on the AU indexes



and extract features that aid ME classification efficiently.

### 2.5.2 Facial Occlusion Detection

In the past literature, Ghiasi et al. (2015) combined an effective part-based model with a binary segmentation technique to precisely identify landmarks and segment out the visible section of the face. Whereas Huang et al. (2022) jointly adopted three network components to segment and identify feature learning for facial occlusion recognition. The two networks, occlusion prediction (OP) and channel refinement (CR) modules, together generate the occlusion mask. The feature purification (FP) network generates occlusion-free, discriminative facial features. Also, Din et al. (2020) proposed a Generative Adversarial Network (GAN) network to generate occlusion-free (non-face object) images. Lee et al. (2020) proposed a face manipulation framework named MaskGAN to enable diverse and interactive face manipulations. The authors developed semantic masks to generate flexible geometric-level face manipulations with fidelity preservation.

Song et al. (2019) proposed a pairwise differential siamese network (PDSN) to construct correspondence between occluded facial blocks and distorted feature elements. Whereas, Opitz et al. (2016) introduced a grid loss layer into sub-blocks of the convolution layer to lower the error rate and provide good performance detecting occluded entities. Loey et al. (2021) proposed a hybrid model using a deep learning model (Resnet 50) and classical machine learning algorithms for facial mask detection. The SVM classifier chosen as the machine learning classifier gave the best results among other classifiers. To detect and segment occluded regions, Chen et al. (2018) proposed a face detector named Adversarial Occlusion-aware Face Detector (AOFD) that gave superior results for both occluded and non-occluded face data. In contrast to previous works, this work focuses on detecting facial occlusions using the Xcep-RA mechanism to further help in efficient feature selection and recognition tasks.

### 2.5.3 Posed Expression Recognition

Ensemble classifier combined the decisions from the multiple classifiers instead of relying on a single classifier decision (Malmasi and Dras 2018). Thus, it helps in

improving the overall accuracy of the system through enhanced decision-making. The ensemble approach proved efficient in various studies [Álvarez et al. \(2016\)](#); [Rao et al. \(2019\)](#); [Sakkis et al. \(2001\)](#) and gave the best accurate prediction. [Wen et al. \(2017\)](#) proposed an ensemble CNN architecture and fused the probabilities of these CNN architectures using the probability-based fusion method. [Yu and Zhang \(2015\)](#) utilized three state-of-the-art face detector modules, ensemble all three face detectors to improve the face detection, and followed the ensemble of various Deep Convolutional Neural Network (DCNN) with randomized initialization for classification of FER. An ensemble classifier based on the Dynamic Weight Majority Voting (DWMV) mechanism with an incremental learning property is proposed by [Zia et al. \(2018\)](#) to learn various incoming expression patterns from images belonging to new expression classes. The combination of SURF with DWMV showed superior performance for FER.

According to [Pramerdorfer and Kampel \(2016\)](#), approaches like data augmentation and ensemble voting improve generalization performance. Hence, they proposed an ensemble of CNN architectures (VGG, Inception, ResNet) without utilizing additional training data or facial registration. This approach became a state-of-art method compared to previous CNN-based FER architectures. In contrast, an ensemble model is proposed by [Kuang et al. \(2016\)](#) with three distinct structured CNN subnets trained separately. The combination of all three ensemble subnets provided better performance results on FER 2013 dataset and obtained 5th rank in the competitions. Also, [Fan et al. \(2018\)](#) proposed a Multi-Region Ensemble CNN (MRE-CNN) framework to detect the contribution of three different sub-regions of the human face. This framework rendered a remarkable performance by assigning the weights to these three networks and combining their final predictions. Finally, an ensemble of Deep CNN's with four different ensemble strategies like a seed, preprocessing, pretraining, and bagging is proposed by [Renda et al. \(2019\)](#) to recognize facial expressions efficiently. The authors have also performed an extensive investigation on various aspects of ensemble generation and focused on the factors which influence classification accuracy. In contrast to the previous approach, this work focuses on

## 2.5. Summary of research works on FER, based on posed, non-posed expressions, facial occlusions and engagement recognition

---

saving the best model weights and extracting their intermediate features, feed them to the base classifiers of the ensemble model to recognize the posed facial expressions and improve the results efficiently.

### 2.5.4 Facial Engagement Analysis

In the past literatures, [Abedi and Khan \(2021b\)](#) combined ResNet and Temporal Convolutional Network (TCN) to extract spatial and temporal variations. The authors overcame the issue of minority-level classification using weighted cross-entropy loss. Similarly, [Liao et al. \(2021\)](#) proposed a Deep Facial Spatio-Temporal Network (DFSTN) that takes a combination of two networks to extract spatial and temporal features for engagement prediction. The fused features aided in getting a fine-grained engaged state and improved the performance of engagement prediction. [Zhu et al. \(2020a\)](#) proposed attention-based Gated Recurrent Unit (GRU), hybrid network for engagement level prediction. The attention-based approach with the deep network aided in extracting important information from the frames.

Also, [Shen et al. \(2021\)](#) proposed a framework based on domain adaptation to assess emotional changes of learners based on facial expressions in a MOOC scenario. Attention mechanism like Squeeze-and Excitation (SE) blocks and Deep Adaptation Network, namely SE-DAN, is proposed to assess learning status in MOOC scenario. Whereas, [Bhardwaj et al. \(2021\)](#) proposed a framework for the detection of engagement and included emotion detection. The haar cascade classifier and CNN model were employed to predict the focus probability. The CNN model was used to calculate Mean Engagement Scores (MES) and classify engagement. [Abedi and Khan \(2021a\)](#) utilized affect states, valence and arousal values, latent affective features, and behavioral features for the measurement of person-oriented engagement in videos. Both frame level and video level analysis were performed. Similarly, [Hao et al. \(2019\)](#) proposed a Weighted Single RGB-stream Inflated 3D Convolutional Network to recognize student engagement automatically. In contrast to previous work, this work focuses on extracting OpenFace toolkit features and Convolutional Recurrent Neural Network (CRNN) features to efficiently classify engagement levels. The fine-grained

## 2. Literature Review

---

features extracted from OpenFace and CRNN are fed to the deep learning model FEA-Net, which helps to improve the discriminative ability of learned features for engagement recognition.

Tables [2.5](#) gives the summary of research works on non-posed expression recognition. Table [2.6](#) gives the summary of research works on facial occlusion. Table [2.7](#) gives the summary of research works on posed expression recognition. Table [2.8](#) gives the summary of research works on facial engagement recognition.

Table 2.5: Summary of research works on non-posed expression recognition. (Note: Listed only some relevant articles considered for the comparison of our objective).

Reference	Datasets	HF	AF	HF + AF	Type of Data	Approaches
Zong et al. (2018)	CASMEII, SMIC			✓	D	Domain Regeneration in the original Feature Space with unchanged Source domain (DRFS-S), Domain Regeneration in the original Feature Space with unchanged Target domain (DRFS-T), Domain Regeneration in the Label Space (DRLS)
Kumar et al. (2019)	SMIC, SAMP, CASME II		✓		D	CNN, Eulerian Motion Magnification (EMM)
Peng et al. (2019)	CASMEII, SAMP, SMIC		✓		D	Apex-Time Network (ATNet) [Two-stream neural network to extract spatial and temporal features using CNN and LSTM architectures.]
Xie et al. (2020)	CASMEII, SAMP, SMIC		✓		D	AU-assisted Graph Attention Convolutional Network (AU-GACN), AU Intensity Controllable Generative Adversarial Nets (AU-ICGAN)
Zhang et al. (2021)	CASMEII, SMIC		✓		D	Intradomain Structure Domain Adaptation (IDSDA)
Takalkar et al. (2021)	CASME, CASMEII, CAS(ME) <sup>2</sup> , SAMP		✓		S	Local and Global Attention Network (LGAttNet)
Yanliang et al. (2021)	SMIC, CASMEII	✓			D	Subspace Learning and Joint Distribution Adaptation (SLJDA)

\* HF–Handcrafted Features, AF–Auto-Extracted Features, S–Static, D–Dynamic

Table 2.6: Summary of research works on facial occlusion detection. (Note: Listed only some relevant articles considered for the comparison of our objective).

Reference	Datasets	HF	AF	HF + AF	Type of Data	Approaches
Huang et al. (2021)	Webface-OCC, LFW, Celebrities in Frontal Profile-Frontal Profile (CFP-FP), AgeDB-30, Real-World Masked Face Recognition Dataset (RMFRD)		✓		S	ArcFace method for face recognition
Loey et al. (2021)	Real-World Masked Face Dataset (RMFD), Simulated Masked Face Dataset (SMFD), LFW			✓	S	ResNet-50, SVM, decision trees, and ensemble algorithms are utilized for classification.
Song et al. (2019)	LFW, MegaFace Challenge, AR face database		✓		S	Pairwise Differential Siamese Network (PDSN)
Wan and Chen (2017)	LFW, AR face database		✓		S	Mask-Maxout, Mask-ResNet
Wang et al. (2023)	Masked Face Detection Dataset (MFDD), RMFRD, Synthetic Masked Face Recognition Dataset (SMFRD)		✓		S	ArcFace method

\* HF–Handcrafted Features, AF–Auto-Extracted Features, S–Static, D–Dynamic

Table 2.7: Summary of research works on posed expression recognition. (Note: Listed only some relevant articles considered for the comparison of our objective).

Dataset	Reference	HF	AF	HF + AF	Type of Data	Approaches
Oulu-CASIA	Jung et al. (2015)			✓	D	Deep Temporal Appearance-Geometry Network (DTAGN)
	Liu et al. (2016)			✓	D	Discriminative Expressionlet (Dis-ExpLet)
	Tang et al. (2018)			✓	D	Geometric features, ANN, CNN (Fusion of Differential Geometric Fusion Network (DGFN) and Deep Facial Sequential Network (DFSN) into DFSN-I using Learning Propagation.)
	Zhang et al. (2017)			✓	D	Part-based Hierarchical Bidirectional Recurrent Neural Network (PHRNN) and Multi-Signal Convolutional Neural Network (MSCNN)
	Cugu et al. (2019)		✓		S	MicroExpNet using Knowledge Distillation (Inception_V3 (basic network))
	Zhao et al. (2016)		✓		S	Peak Piloted Deep Network (PPDN) (GoogLeNet (basic network))
	Yang et al. (2018)		✓		S	De-expression Residue Learning (DeRL) (conditional Generative Adversarial Network (cGAN) is utilized as a generative model)
	Ding et al. (2017)		✓		S	FaceNet2ExpNet (FaceNet (VGG-16 (basic network)); ExpressionNet (CNN followed with Rectified Linear Unit (ReLU) and Max pooling layer (basic network)))
RaFD	Wenyun et al. (2018)		✓		S	18-layered Convolutional Deconvolutional Networks (Conv-Deconv)
	Fathallah et al. (2017)		✓		S	CNN (VGG network (basic network))
	Sun et al. (2019)		✓		S	Multi-channel Deep Spatial-Temporal feature Fusion neural Network (MDSTFN) (Inception Network (basic network))
	González-Hernández et al. (2018)		✓		S	CNN

\* HF–Handcrafted Features, AF–Auto-Extracted Features, S–Static, D–Dynamic

Dataset	Reference	HF	AF	HF + AF	Type of Data	Approaches
RaFD	Happy et al. (2019)		✓		S	Weakly supervised learning (CNN)
	Yaddaden et al. (2018)	✓			S	Hybrid based approach (Geometric features (fiducial points), appearance features (Discrete Wavelet Transform)), SVM
	Jiang and Jia (2016)	✓			S	Local Discriminative Component Analysis (LDCA)
	Sun et al. (2017)			✓	S	Binarized Neural Networks (BNNs), Binarized Auto-Encoders (BAEs), Multi-scale Dense Local Binary Patterns (MDLBP)
	Carcagnì et al. (2015)	✓			S	HOG, SVM
	Jiang and Jia (2013)		✓		S	Multi-Pose Adaptive Boosting (MP-Adaboost)
	Rao et al. (2015)	✓			S	Speeded-Up Robust Features (SURF) boosting
	Wu and Lin (2018)		✓		S	Weighted Center Regression Adaptive Feature Mapping (W-CR-AFM)
	Shokrani et al. (2014)	✓			S	Pyramid Histogram of Oriented Gradient (PHOG)- K-Nearest Neighbors (K-NN)
	Kurup et al. (2019)			✓	S	Semi-supervised DBN

\* HF–Handcrafted Features, AF–Auto-Extracted Features, S–Static, D–Dynamic



Table 2.8: Summary of research works on facial engagement recognition. (Note: Listed only some relevant articles considered for the comparison of our objective).

Reference	Datasets	HF	AF	HF + AF	Type of Data	Approaches
Abedi and Khan (2021a)	DAiSEE, EmotiW			✓	D	Latent affective, behavioral and affect features with Temporal Convolution Network (TCN)
Abedi and Khan (2021b)	DAiSEE		✓		D	Resnet and TCN
Liao et al. (2021)	DAiSEE		✓		D	Deep Facial Spatiotemporal Network (DFSTN) [Squeeze-and-Excitation ResNet-50 (SE-ResNet-50 (SENet)) , LSTM]
Hao et al. (2019)	DAiSEE		✓		D	Weighted Single RGB-stream Inflated 3D ConvNet (WSRGB-I3D)
Gupta et al. (2016a)	DAiSEE		✓		D	Convolutional 3D (C3D) with Fine Tuning, Long-term Recurrent Convolutional Network (LRCN)
Geng et al. (2019)	DAiSEE		✓		D	C3D with focal loss

\* HF–Handcrafted Features, AF-Auto-Extracted Features, S–Static, D–Dynamic

### 2.6 OVERALL FINDINGS FROM THE LITERATURE REVIEW

It is challenging to track subtle movements of facial muscles due to the availability of limited datasets. Hence, it needs exploiting of powerful features that can characterize facial expressions into emotion categories (Pan et al. 2019). Techniques that work well on datasets collected in a controlled environment may work worse when tested in natural and unconstrained environments (Barros et al. 2017; Li et al. 2018) due to its distinct AU's (Li and Deng 2018b; Li et al. 2017). Comparing the data collected in an unconstrained environment to that of a controlled environment, it contains a diverse set of AUs (Li and Deng 2018b); few AUs are common, while few depend on culture and on the time of context. Figure 2.2 depicts the diversity of AUs in a constrained and unconstrained environment. Consider an example of surprise expression class in a constrained and unconstrained environment. As observed, a few AUs like AU1, AU2, AU5, and AU25 are common among a few individuals, while a few AUs like AU12, AU20, AU26, and AU16 occur and vary between individuals based on context and time. Subtle changes in facial features, co-occurrences between AU's, and the low resolution make it harder for the model to recognize these emotions efficiently. The network that learns subtle facial dynamic patterns from micro-expression can efficiently work well on recognizing posed expressions occurring in a lab environment (Kim et al. 2017). Also, psychologists can further investigate and explore the types of characteristics (AU's differ for each side of the face, there is also variation in intensity) required to classify the real-world data (Yan et al. 2020). Small minute changes in the facial region are difficult to detect. Semi-automated systems are employed in the detection of Micro-Expressions (MEs). Consequently, there is a lack of a fully integrated framework to analyze MEs (Davison et al. 2016).

The structural changes in the regions of facial landmarks like eyebrows, eyes, nose, and mouth are essential and, in many cases, sufficient for AU recognition. The past literature has shown that the efficient categorization of expressions benefits from localization and feature extraction from areas like the eyes, nose, and mouth (Gizatdinova and Surakka 2008; Tian et al. 2000). There are upper face AUs, which predominantly help in the categorization of expressions efficiently. Some upper face

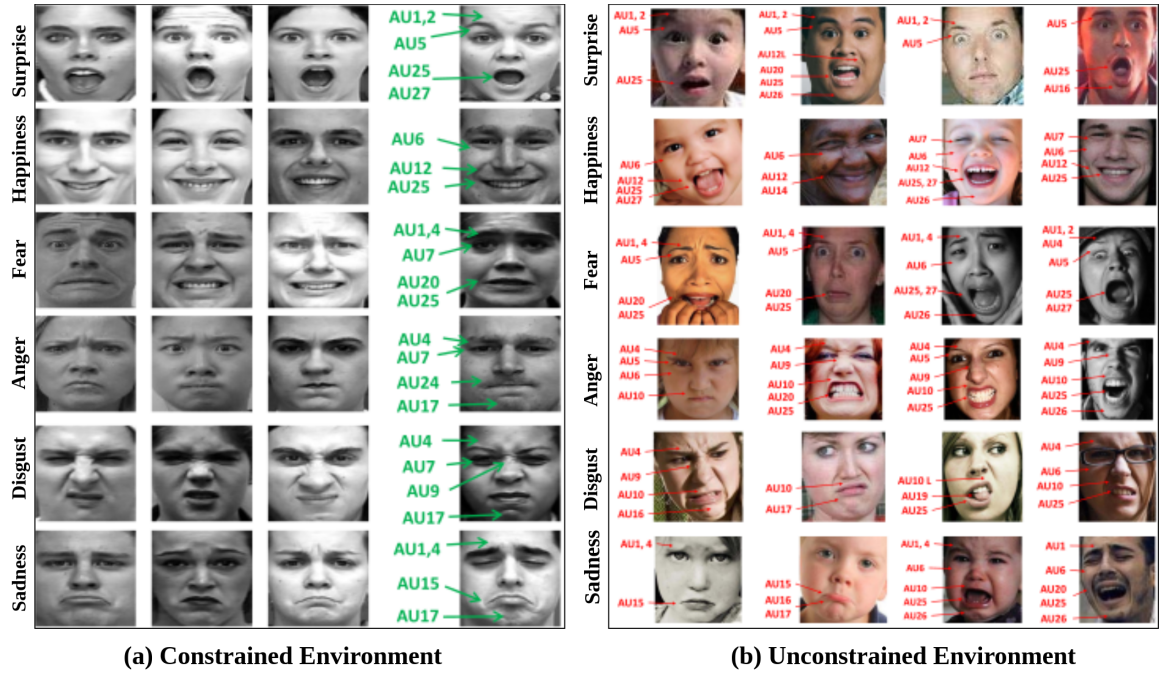


Figure 2.2: The diversity of AUs in constrained and unconstrained environment. The Figure is adapted from [Li and Deng \(2018b\)](#).

AUs that give more information are the AU1-inner portion of the brow raised, AU2-outer portion of the brow is raised, AU4-brows lowered and drawn together, AU5-upper eyelids are raised, AU6-cheeks are raised, and AU7-lower eyelids are raised (lid tightener). Similarly, the lower face AUs that primarily help in the categorization of expressions are AU9-nose wrinkler, AU10-upper lip raiser, AU11-chin raiser, AU12-lip corner depressor, AU14-lips part, AU15-jaw drop, AU16-mouth stretch, AU17-lower lip depressor, AU18-lip pucker, AU20-lip tightener, AU23-lip presser, AU24-nasolabial furrow deepener, AU25-lip corner puller, AU26-lip stretcher, and AU27-dimpler.

Primary emotions are universal. FER system poses a lot of challenges in recognizing emotion classes like sadness, fear, and disgust. There is a need for a robust system that efficiently recognizes these universal emotions in both constraint and unconstraint environments. Confusions arise when classifying fear and sadness, fear and surprise, disgust, and sad emotions ([Li et al. 2018](#)). Expressions like anger, sadness, and fear vary and depend on ethnicity, which makes recognition and capturing of such expressions a challenging task ([Valstar and Pantic 2010](#)). There is no proper

consensus for universal expressions from psychologists on the relationship between patterns of AUs and emotions (Yan et al. 2020)

Various factors like head rotation, pose variations, occlusions, illumination variations, differences in age, gender, culture, and skin tone makes it challenging for training and testing the models in an unconstrained environment. It is challenging to select efficient feature extraction and classification techniques, which work well in these varying conditions (Majumder et al. 2016). Neither conventional nor deep learning approaches are robust to overcome all the challenges in an in-the-wild dataset (Sepas-Moghaddam et al. 2021). Hence, it remains a challenge in the FER system. An ideal FER system needs to be developed, which can handle all these challenges in real-life situations, as these factors cause changes in visual appearance and deteriorate FER systems' performance. Addressing the issue of facial occlusions is trivial (Li et al. 2018), as it varies in their positions and occluders. Also, working on determining specific parameters, for the detection of facial occlusions and proper pre-processing techniques, automatically needs further exploration in the future. Thus, the robust deep learning technique with attention mechanism is of utmost importance to be developed, capable of focusing on unblocked facial patches and perceiving informative features from them to help classify the expressions into intended classes.

### 2.7 RESEARCH GAPS IDENTIFIED

Some of the important research gaps identified from the above literature review are:

- Most of the research work in the literature has focused on basic emotions; Extending the work by identifying learning-centered emotions or affective states non-intrusively is essential.
- Emotion recognition varies depending on the constrained and unconstrained environment, which is a challenging issue that needs to be addressed.
- Recursive selection of important features and important classifier is required to get accurate recognition of emotions.
- Recognition of universal expressions is challenging and needs further exploration,

as there arises confusion between fear and sadness, fear and surprise, disgust and sad emotions during classification.

- Recognition of micro-expressions is challenging and needs further exploration. A precise technique or semi-automated system is required to record minute changes in the facial region and get satisfactory results.
- Elicitation of micro-expression data are challenging as compared to macro-expression data.
- Recognition of micro and macro facial expressions from partially occluded faces are challenging. Detection of occlusions is difficult as it varies in positions and occluders. Robust deep learning approach with an attention mechanism, capable of focusing on unblocked facial patches and perceiving informative features needs to be adapted.
- Head rotation, face pose, illumination variation, occlusion, etc. are the attributes that increase the complexity of recognition of spontaneous expressions in practical applications. It will also lead to data loss, which needs to be addressed appropriately.
- In-the-wild data still remains a open challenge. Techniques that work well on datasets collected in a controlled environment may work worse when tested in natural and unconstrained environments. Subtle changes in features, co-occurrences between AU's, and low resolution make it harder for the model to recognize these emotions efficiently.
- Accurate localization of facial landmarks is an important building block for identification & analysis of facial expressions that need to be considered for many applications.
- Classifying facial expression by having multi-model fusion would probably increase the accuracy of the system, which needs to be addressed.
- Emotions change from person-to-person, i.e., with respect to age, gender, culture, and ethnicity. Emotion recognition concerning the age group is not found in the

literature. Assessment of the affective states varies greatly from person to person, so a personalized affective analysis can achieve better performance and improve usability.

- Prediction of emotions can be improved further when time and context are considered, which need to be addressed.
- Generalization of model or system for all subjects for predicting affective states is a challenging issue that needs to be further addressed.

### 2.8 PROBLEM STATEMENT

The primary aim of this work is design and development of Artificial Intelligence based techniques for facial emotion recognition from posed and non-posed facial expressions.

### 2.9 PROBLEM DESCRIPTION

The defined problem is further divided into five tasks and elaborated with little insight. The research aims to recognize posed and non-posed expressions from facial images using neural network architectures. To provide a localization technique that detects facial occlusions that arises in unconstrained environment settings to aid the efficacy of feature selection and the proper recognition processes. Finally, assessing engagement levels in Massive Open Online Courses (MOOC) scenarios to overcome educational problems like reducing dropout rates and addressing low achievements of students in academics.

The objectives of the work are:

1. Design and development of an efficient technique to extract relevant features from images for efficient emotion classification.
2. Propose a localization mechanism to address the issue of facial occlusions effectively.
3. Detection of posed expression using deep neural network.
4. Develop a novel ensemble model for the detection of non-posed expression from visual cues.

5. Automation and systematic analysis of engagement levels from visual cues.

## 2.10 DATASETS CONSIDERED FOR THIS THESIS

This section familiarizes the datasets used in this research work to perform experimental analysis. Table 2.9 gives the details of the database considered for each objective.

1. Oulu-CASIA: This facial expression database is a posed dataset (Zhao et al. 2011). It includes 480 image sequences elicited from 80 subjects in six different background settings using three illumination conditions: normal, weak, and dark. The cameras like Near-Infrared (NIR) and VISual (VIS) were used to capture the same facial expressions elicited by subjects. Each image sequence in the database begins with a neutral expression and ends with peak emotion labels. Each image's pixel resolution is 320\*240. This dataset includes basic emotions like anger, disgust, fear, happiness, sadness, and surprise.
2. Radboud Faces Database (RaFD): The RaFD database (Langner et al. 2010; Shokrani et al. 2014) is a posed dataset which contains images from 67 subjects collected using five camera angles and has five pose degrees 0, 45, 90, 135, 180 with three gaze directions (frontal, left and right views). The dataset includes expressions like anger, disgust, fear, happiness, sadness, surprise, contempt, and neutral. A total of 8040 images are present in this database, and each image pixel resolution is 681\*1024.
3. Chinese Academy of Sciences Micro-Expression II (CASME II): The CASME-II dataset is an improved version of CASME that uses video clips to elicit the microexpressions of its subjects (Yan et al. 2014). Each sample's onset, apex, and offset frames, including the AUs label, are annotated. The recordings from 26 subjects were obtained, creating 255 micro-expression samples with a resolution of  $640 \times 480$  and a sampling rate of 200 FPS. Four Light-Emitting Diodes (LEDs) are exploited for brightness compensation. The dataset includes seven expression categories: disgust, fear, happiness, repression, sadness, surprise, and others.

4. Spontaneous Actions and Micro-Movements (SAMM): SAMM is a high-resolution spontaneous dataset (Davison et al. 2016). The micro-movements of 32 participants are induced spontaneously with the widest range of demographics. The mean age of participants was 33.24 years, with an even gender distribution of 16 male and female participants. The dataset was designed to be as diverse as possible to elicit a wide range of emotional responses.
5. Spontaneous Micro-expression Corpus (SMIC): Three cameras are used to record the SMIC dataset: High-Speed (HS), VIS, and NIR cameras (Li et al. 2013; Pfister et al. 2011). The HS data is recorded with a sampling rate of 100 FPS, and it includes 164 micro-expression samples. Whereas VIS and NIR data are recorded with a 25 FPS using ordinary and near-infrared cameras, respectively, and both VIS and NIR data contain 71 micro-expression samples individually. The resolution of video sequences is  $640 \times 480$ , and most of the participants are from an Asian ethnicity. The dataset includes three categories of expressions: positive, negative, and surprise.
6. Webface-OCC: It is a simulated occluded dataset with 804,704 face images collected from 10,575 subjects. The dataset is based on the CASIA-webface dataset (Yi et al. 2014), which combines unmasked faces with those with simulated occlusions. The occluded appearance on the actual faces is achieved by concealing a variety of occluded objects such as glasses or masks, each with a unique texture and color (Mare et al. 2021).
7. Labeled Faces in the Wild (LFW): LFW is a benchmark dataset utilized for public face recognition, and the data is collected in an unconstrained environment (Learned-Miller et al. 2016). In the LFW Simulated Masked Face Dataset, there are 70 masked face images of 48 people in the test set and 13027 masked face images of 5713 people in the train set.
8. Real-World Masked Face Dataset (RMFD): RMFD dataset contains a diverse set of real masked faces, consisting of 5000 masked and 90,000 unmasked images



of the face. This dataset is considered the world’s largest real-world masked face dataset, and the face images are crawled from massive Internet resources (Wang et al. 2023).

9. Dataset for Affective States in E-learning Environments (DAiSEE): DAiSEE consists of 9068 video snippets of 10 seconds long, which are recorded from 112 individuals (Gupta et al. 2016a). The frames were captured at 30 frames per second with a full High Definition (HD) webcam at a resolution of 1920×1080 pixels. The dataset contains four affective states: bored, frustrated, confused, and engaged.

Table 2.9: The details of datasets considered in this thesis.

Sl.No.	Name of the dataset	Objective
1	Oulu-CASIA	Objective-3
2	Radboud Faces Database (RaFD)	Objective-3
3	Chinese Academy of Sciences Micro-Expression II (CASME II)	Objectives-1 & 4
4	Spontaneous Actions and Micro-Movements (SAMM)	Objectives-1 & 4
5	Spontaneous Micro-expression Corpus (SMIC)	Objectives-1 & 4
6	Webface-OCC	Objective-2
7	Labeled Faces in the Wild (LFW)	Objective-2
8	Real-World Masked Face Dataset (RMFD)	Objective-2
9	Dataset for Affective States in E-learning Environments (DAiSEE)	Objective-5

## 2.11 SUMMARY

This chapter comprehensively reviews and summarizes the technologies and existing problems in this area. Existing scientific issues, real-world applications, and future directions presented in this chapter aid the researchers explore the field efficiently. Even though numerous works have been carried out in the field of FER, systems do suffer from some drawbacks that need some efficient solution and further exploration shortly. There is a requirement to build an adequate system that is robust in constrained and unconstrained environments and aims to achieve accurate recognition as these methods will be exceedingly helpful in real-life scenarios. This chapter briefly highlights the

## *2. Literature Review*

---

information on the databases used to conduct the experiments for all objectives. It also includes the research gaps discovered through the literature review, problem statement, and objectives. Chapter 3 discusses on the approach proposed to solve the first and fourth objectives, the recognition of non-posed expressions.

## CHAPTER 3

# FEATURE EXTRACTION AND CLASSIFICATION OF NON-POSED EXPRESSIONS

This chapter briefly explains the methods for extracting features and identifying non-posed expressions. The facial regions are segmented depending on the AU description using mesh generation techniques. The segmented regions are used to extract the geometric and texture properties. Further, non-posed expressions are categorized using ensemble techniques. This work has created a Micro-Expression Recognition (MER) system utilizing AU indexes to extract features and analyze Micro-Expressions (MEs).

### 3.1 INTRODUCTION

MEs have a brief duration and a minute movement amplitude. Detecting and recognizing MEs has been crucial in criminal investigation, suicide intervention, and defense. ME analysis has benefited from the AU-based study (Dong et al. 2022). This work focuses on AU analysis and considers each AUs contribution to the analysis of expression, and every AU has additional aid for recognition of each facial expression (Zhi et al. 2021). The scientists have agreed that facial deformations convey various efficient information for social communication (Barrett et al. 2019). Locating specific FACS provides better discrimination between ME classes (Liong et al. 2018). Also, there exist strong dependencies between the facial expressions and the AUs, which can aid in guiding the learning process of the model (He et al. 2021).

### 3. Feature Extraction and Classification of Non-posed Expressions

---

There are two categories of expressions like macro and MEs; both differ in relative duration and intensity. MEs are spontaneous expressions with facial muscle movements which last for not more than 200ms (Xie et al. 2020). ME reflects the true feelings of humans; thus, the recognition of such expressions is helpful to apply in real-world applications. MER plays an essential role in presenting people's real emotions and is helpful in high-risk situations and various real-world applications. ME has a short duration, low intensity, and fragmental facial action units occurring in only some part of the face, thus making it difficult to recognize such expressions with the naked eye (Oh et al. 2018; Wang et al. 2014). This work proposes a novel MER system by using Delaunay Triangulation (DT) and Voronoi Diagram (VD) approach based on AUs description. Distinct AUs are modeled, and their corresponding geometric and texture features are extracted and integrated, which help in the recognition of minute changes over the face. Instead of building a new ME dataset, building an optimal solution by processing the ME samples into distinct features and using the machine learning classifier for classification appears to be a more efficient process (Thi Thu Nguyen et al. 2021). Inspired by this issue, this work utilizes the DT and VD to segment the face and extract the geometric and texture features based on the AUs from neutral and peak frames for efficient ME recognition. Effective analysis of facial expression heavily depends on accurate representation of facial features (Happy and Routray 2014). DT is an efficient and unique approach to segment the face and also for spotting the ME's (Oh et al. 2018) thus giving freedom to define the Region of Interest (ROI) of the face. Usage of ROI is effective compared to the entire face, and extraction of features from those ROI's is required to analyze ME's. This work utilizes the DT approach to calculate the whole face's geometric features. Also, Voronoi Tessellations or VD extract ROI and LBP features from them. These features are appended and fed into the classifier (ensemble model) for classification.

Instead of using a whole video sequence, the peak frame can easily express emotion as well provide significant performance (Khor et al. 2019). Onset and apex frames are sufficient to be processed for encoding recognition of ME features (Liong et al. 2018). Thus retain the computational simplicity by avoiding the processing of entire

video sequences and eliminating redundant information by eradicating repetitive frames with minute changes (Liong and Wong 2017). In this experiment, training and testing samples are from different databases. Hence, there is a significant difference in feature distribution, leading to a decreased performance for MER.

This study aims to exploit geometric and texture features structured by DT and VD, respectively, to select relevant features for MER. The overview of the proposed method is illustrated in Figure 3.1. In this experiment, the first stage is image acquisition, and an apex-based strategy is used as a frame selection method that considers two frames onset and apex. Next, the facial region is detected using facial landmarks. Locations of facial points are used for defining the facial regions. As the facial expression evolves, the location of facial landmarks helps in extracting the shape and movement of facial features. These facial landmarks are mapped into corresponding AUs based on the facial movements. Thus, it helps in extracting the ROI from the whole facial region. In the second stage, the proposed method is further partitioned into two significant sub-parts. In the first part, using the DT procedure based on AU indexes, the ROI is obtained, then the geometric features from each individual AU region are extracted. The second part extracts LBP texture features from ROI obtained using the VD based on AUs indexes. Thus, ROI is obtained using DT and Voronoi Tessellations instead of considering the whole facial features. Every individual AU and its corresponding geometric and texture features play an important role in classifying MEs. Finally, the feature difference between the onset and peak frames geometric and texture features are clubbed into a vector and fed into an 'N' number of the ensemble machine learning classifiers for the classification in the third stage. The flow of the proposed methodology is discussed briefly in subsection 3.3 further.

The contributions of the chapter are :

- Identification and mapping of the landmark indexes based on AUs and finding the relationship between the AUs and MEs.
- DT and VD properties are utilized to segment ROI based on AU indexes. The ROI-based feature extraction reduced the feature dimension and improved the

### 3. Feature Extraction and Classification of Non-posed Expressions

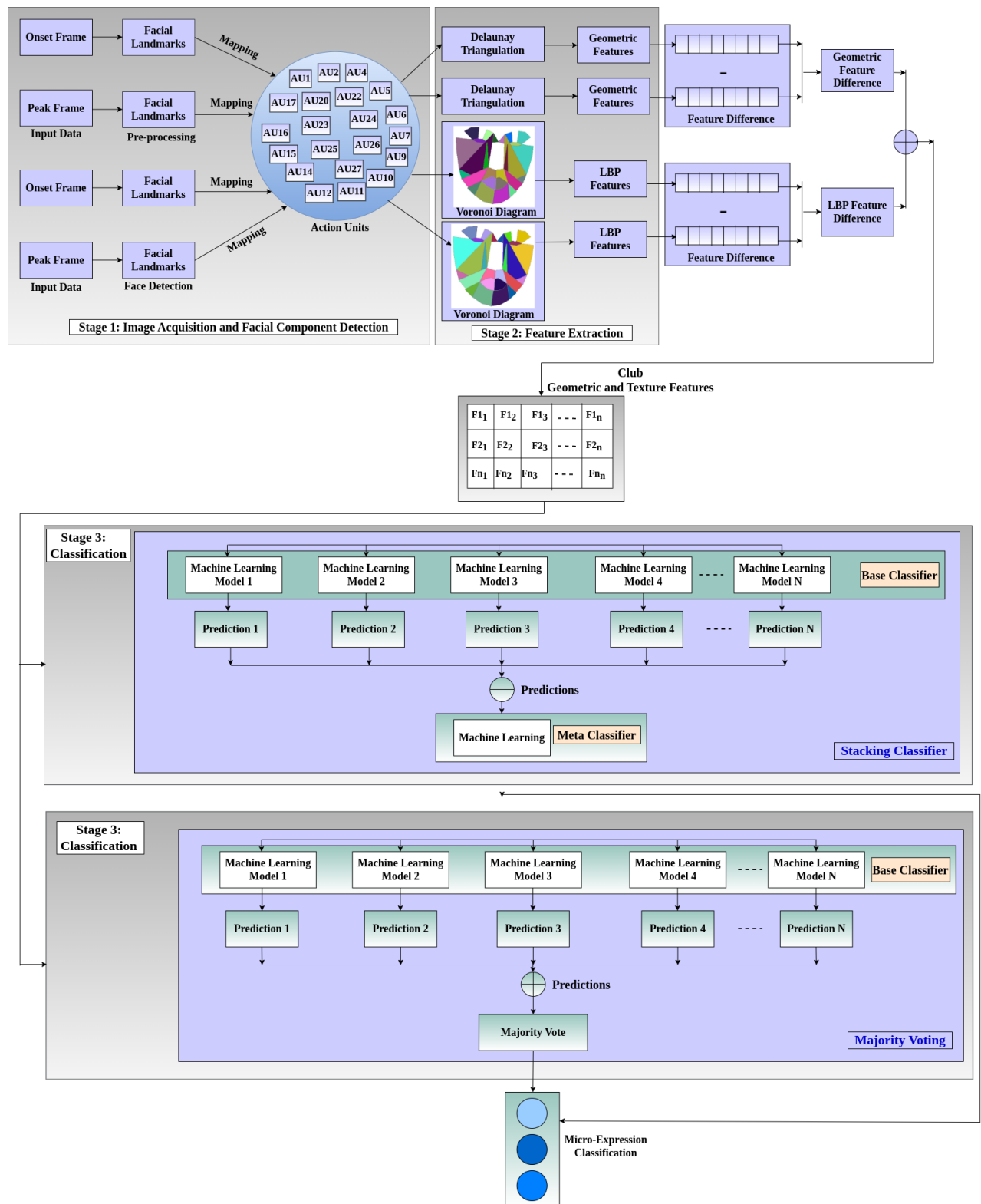


Figure 3.1: Facial segmentation and feature extraction system for MER.

MER performance.

- The proposed work employs AU indexes to solve the problem of DT. Thus, retaining the important triangles helpful for the detection of MEs.

The chapter is organized as follows: Section 3.1 introduces the feature extraction technique. Section 3.2 briefs preliminary concepts associated with DT and VD. Section 3.3 describes the proposed feature extraction technique along with its procedure. Section 3.4 details the experimental setup and analysis of results, and Section 3.5 summarizes the proposed method and its significance.

## 3.2 PRELIMINARIES

The DT and the VD have been extensively studied and applied in a variety of fields (Jiang and Jiang 2019; Khodadoust and Khodadoust 2017). DT is a mesh generation technique invented by Delaunay with the capability for autonomous operation. The compactness of the produced mesh is a criterion for an automatic mesh generating algorithm. According to Euler-Poincare, the association between the vertices, edges, and faces should be precise. At the end of the creation process, there should not be any loose vertices, gangling edges, or faces. The DT technique is employed for filling a volume  $V$  with three-dimensional triangles (i.e., tetrahedra is one of the Delaunay-based techniques). Given a finite collection of 'P' points, a triangulation of 'P' is a simplicial complex that tessellates the convex hull of 'P' and whose vertices belong to 'P'. Also, multiple triangulations are created using the same set of points.

The DT is the most effective technique and has good properties like dynamic generation and reliable geometry. The triangulation construction is stable because, in the original DT, insertion and deletion operation concerning one vertex only require updating of the local region. The properties of DT significantly reduce time expenditures when frequent editing procedures are necessary. Also, the minor angle among the triangles is maximized in the triangular mesh to avoid a skinny triangle around the empty circle property, thus ensuring that the circumcircle of any triangle has no vertex points in its interior. Using the points deduced from image 'i', the triangulation 'T' can be constructed where each point defines a vertex  $v_i$  in  $V$  such that

$V = \{v_i; i = 1, 2, \dots, n\}$ , and each triangle comprised of three edges in  $E$  satisfies the DT's empty circle property (Jiang and Jiang 2019).

The DT (Cheddad et al. 2008) is the dual tessellation of the VD also known as Voronoi Tessellation for the sampling solutions  $S_i, i=1,2,\dots,N$ . The VD is a collection of polygons represented as  $P_i$ , where each of these polygons is centered at  $S_i$ , such that it contains all the points that are nearest to  $S_i$  than any other data point (Bebis et al. 1999). Also, if these polygons have a common edge, the DT is produced by drawing the line segments between Voronoi vertices. Once the VD for a set of points has been generated, DT is as simple as joining any two sites whose Voronoi polygons share an edge. The VD is constructed given a collection  $\{P\}$  of  $m$  identical random points in 2D volume. There is a local sub-region  $V_i$  connected with each point  $P_i \in \{P\}$  such that  $V_i = \{X : \|X - P_i\| < X : \|X - P_j\|, \text{ for all } j \neq i\}$ . A group of sub-regions,  $V = \{V_i, i = 1, 2, \dots, m\}$  is described as a Voronoi Tessellation of the volume. According to the criteria, the 2D volume space is randomly partitioned into a compact set of the convex hull, which is associated with randomly produced  $m$  interior points. Each sub-region of  $V_i$  can be represented as the convex intersection of the open half-spaces (lines or planes) separating the points  $P_i$  and  $P_j$  ( $j \neq i$ ). The Voronoi Tessellation of the convex hull also satisfies the Euler-Poincare relation, and that each internal edge or face of  $V_i$  is shared by only two convex hulls.

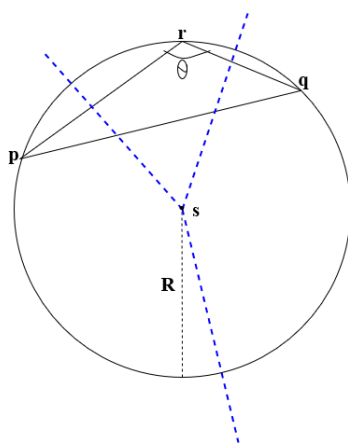


Figure 3.2: Delaunay Triangulation (DT) construction (Cheddad et al. 2008).

For a better understanding of DT and VD properties, the construction is depicted in



Figure 3.2, where P be a set to be more explicit. The empty circle property holds, if there is no other point P inside the circle bounded by the triangle pqr, and its center lie on a Voronoi vertex, the triangle pqr is called a DT. The center (S) of the circle bounded by pqr is a Voronoi vertex.

### 3.3 PROPOSED FEATURE EXTRACTION AND CLASSIFICATION TECHNIQUE

The overall workflow of the proposed framework is illustrated in Figure 3.3. Face detection and facial component detection are carried out in the first stage. The 68 facial landmark indexes are retrieved from the input image and mapped using AUs. The point set composed of landmark indexes based on AU information is input into a triangulator in the second stage as a first portion, and triangulation is obtained (segmented region). Additionally, the triangulated regions' geometric features are extracted, and the feature difference is calculated. To create a voronoi-based segmented region, the point set comprising landmark indexes based on AU information from the second portion in the second stage is passed into a tessellator. Additionally, the segmented region's texture characteristics are extracted, and the feature difference is calculated. Further, both features are clubbed and processed to the third stage. In the third stage, stacking classifier and majority voting is utilized individually to perform the classification task. A brief explanation of each workflow stage is provided further.

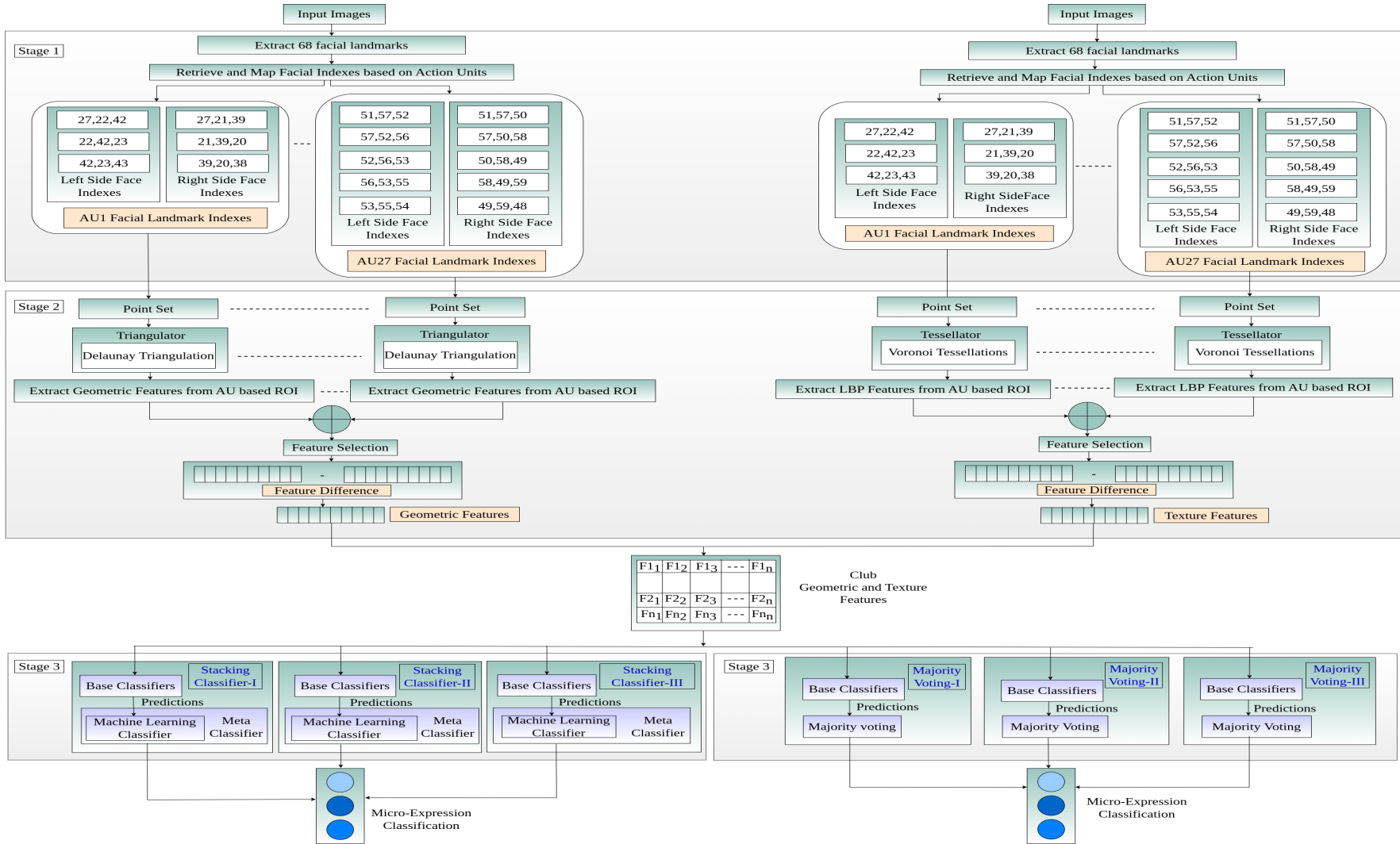


Figure 3.3: Detailed procedure outlining the flow of proposed methodology.

### 3.3.1 Stage 1: Face Detection

- 1 Image Acquisition: The apex-based strategy is utilized in this experiment, where onset and peak frames are extracted from the datasets instead of considering the entire video sequences and eradicating repetitive frames. As suggested by [Kun-Hong et al. \(2021\)](#), the apex-based strategy is relatively a new solution that needs to be adapted for carrying out the recognition of MEs. Due to the limitation in the number of images in the ME dataset, offline data augmentation technique is performed like bilateral filtering (with varying parameters like diameter, sigmaColor, and sigmaSpace), average blurring (with varying kernel size=(3,5,7,9)), gaussian blurring (with varying kernel size=(3,5,7,9)), median blurring (with varying kernel size=(3,5,7,9)). As a result, data augmentation is performed on the training set, overcoming the data limitation.
- 2 Facial Component Detection: Facial landmark aid the identification of facial components and is obtained using the shape predictor Dlib, as illustrated in Figure [3.4](#). The shape predictor locates 68 landmark points on the facial image based on the previously trained model ([Munasinghe 2018](#); [Shahar and Hel-Or 2019](#)). In this experiment, facial landmarks are obtained for both onset and peak frames independently. The landmark indexes are recorded into Comma-Separated Values (CSV) files, which are subsequently used to extract further characteristics.

MEs are based on the combination of distinct facial regions instead of considering only one area. This motivated us to select the most significant facial areas based on AUs indexes. The extracted facial landmark points are mapped into distinct AUs based on facial components and their muscle movements. Later, these indexes are used as a seed for constructing DT and Voronoi Tessellations for feature extraction. The FACS system designed by Paul Ekman helps to describe facial expressions by defining a set of atomic facial muscle actions known as AUs ([Liu et al. 2019](#)). AUs play an essential role in finding the minute changes in facial movements and extracting meaningful feature information ([Xie et al. 2020](#)). The presence of AUs and their combination

### 3. Feature Extraction and Classification of Non-posed Expressions

---

together help discover the differences in muscle movements and further help to map into respective emotion classes. The facial actions describe the local variations on the face (Zhi et al. 2021). Each AU is coded based on the contraction of a single or group of muscles. Under distinct facial expressions, certain AUs show strong relationships (Xie et al. 2020). FACS is an anatomical system that encodes various facial movements by combining distinct basic AUs, thus making the categories of emotions much wider.



Figure 3.4: Facial landmarks obtained from Dlib function (Shahar and Hel-Or 2019).

Since more attention needs to be paid to the research addressing the relationship between facial actions and emotion labels (Zhi et al. 2020). This work aims to examine the subtle facial feature difference that exists amongst the facial action units, which could aid in classifying the expressions efficiently. Each AU contributes differently to the classification of facial expressions. Since it is tedious and challenging to generate facial representations for vast quantities of AUs (Zhi et al. 2020), this work focuses on analyzing the AUs and finding the representations for AUs of only important facial regions that would strongly help for classification.

First, the AUs are recognized, facial features are extracted from those AU regions. The MEs are deduced from the identified AUs and their characteristics. Thus, AUs are regarded as building blocks of facial expressions in the FER system. This study integrates AU relevant face regions based on matching

landmark indexes (68 landmark indexes starting from index 0) as illustrated in Figure 3.4. Each facial feature component and its related landmark indexes are listed in Table 3.1. These landmark indexes are mapped separately into various AUs based on the facial muscle movements as listed in Table 3.2. Later these indexes are used to partition the facial region depending on the AUs using DT and Voronoi Tessellation approaches. Each of the AUs analyzed contributes to the creation of an MER system. Further, the geometric and texture features are retrieved from the segmented ROI, as illustrated in further subsections.

Table 3.1: Landmark indexes considered for evaluation of the facial feature components.

Facial Feature Components	Facial Landmark Indexes
Mouth	(48, 67)
Mouth_outline_points	(48, 60)
Mouth_inner_points	(61, 67)
Right_Eyebrow	(17, 21)
Left_Eyebrow	(22, 26)
Right_Eye	(36, 41)
Left_Eye	(42, 47)
Nose	(27, 35)
Jaw	(0, 16)

### 3.3.2 Stage 2: Feature Extraction

Extraction of powerful features and building an efficient classifier are essential approaches for traditional FER systems. In this approach, feature extraction and classification task are performed individually. By representing features adequately, there can be an increase in the recognition system's efficiency, thus minimizing the within-class variations and maximizing the between-class variations. There are two types of feature extraction techniques geometric-based and appearance-based extraction (Majumder et al. 2016). Geometric features use the location of facial features or the shape of the facial components. Even though geometric features are noise sensitive, they need precise detection of landmarks and alignment information. Still, they prove to be efficient in providing accurate results in recognition of facial expressions (Ghazouani 2021). The appearance feature describes the texture and

### 3. Feature Extraction and Classification of Non-posed Expressions

Table 3.2: Mapping of action units based on landmark indexes considered for the proposed methodology.

Action Units	Description	Left Side Landmark indexes considered for each AU's	Right Side Landmark indexes considered for each AU's
AU1	Inner Brow Raiser	22,23,27,42,43	20,21,27,38,39
AU2	Outer Brow Raiser	24,25,26,44,45	17,18,19,36,37
AU4	Brow Lowerer	22,27,42	21,27,39
AU5	Upper Lid Raiser	43,44,46,47	37,38,40,41
AU6	Cheek Raiser	13,14,15,16,42,45,46,47	0,1,2,3,36,39,40,41
AU7	Lid Tightener	42,43,44,45,46,47	36,37,38,39,40,41
AU9	Nose Wrinkler	13,22,35,54	3,21,31,48
AU10	Upper Lip Raiser	34,51,52,54,57	32,48,50,51,57
AU11	Nasolabial Deepener	35,42,54	31,39,48
AU12	Lip Corner Puller	12,13,54	3,4,48
AU14	Dimpler	12,35,54	4,31,48
AU15	Lip Corner Depressor	8,10,12,35,54,57	4,6,8,31,48,57
AU16	Lower Lip Depressor	8,9,10,55,56,57	6,7,8,57,58,59
AU17	Chin Raiser	8,9,10,11,54,55,56,57	5,6,7,8,48,57,58,59
AU20	Lip Stretcher	48,51,54,57	48,51,54,57
AU22	Lip Funneler	35,51,54,57	31,48,51,57
AU23	Lip Tightner	33,35,51,54,57	31,33,48,51,57
AU24	Lips Pressor	51,52,53,54,55,56,57	48,49,50,51,57,58,59
AU25	Lips Part	51,52,53,55,56,57	49,50,51,57,58,59
AU26	Jaw Drop	8,9,10,51,52,53,54	6,7,8,48,49,50,51
AU27	Mouth Stretch	51,52,53,54,55,56,57	48,49,50,51,57,58,59

intensity information, LBP feature (Ojala et al. 1996) is used for the texture analysis as they are tolerant to illumination variation, misalignment error (Oh et al. 2018; Senechal et al. 2014) and computational simplicity. Appearance features contain micro-patterns that helps in providing important information about the facial expressions, but they fail to generalize across distinct subjects (Ghazouani 2021). Both the feature representations like geometric features and appearance have their benefits and drawbacks. The correlation of multiple facial regions is vital for AU detection and can provide more robust features than individual regions (Li et al. 2021). Thus, combining geometric and texture features from multiple AU segmented

regions is essential in recognizing minute changes in the facial region and aid MER. DT and VD are used in this experiment to segment ROI using AUs and extract geometric and texture features from those significant areas and reduce the ineffective information (Liu et al. 2019).

1. Delaunay Triangulation (DT) based Geometric Feature Extraction: ME's can be recognized efficiently from the specific facial region instead of considering the entire face (Lei et al. 2020). Among all facial regions, the mouth and eyebrows contribute a lot for recognition of ME, and the shape features tend to show more contribution towards muscle movements (Lei et al. 2020). Using the DT and specifically choosing ROI using facial feature points to define the region boundaries can reduce irrelevant facial features (Davison et al. 2016). This led to the decision to use DT to efficiently extract ROI and then extract geometric features from these Regions.

The triangulator takes a finite set of landmark points mapped to AUs as an input. These AUs are utilized as seed points to build the DT. In this work, AU indexes are fed to construct the DT, forcing the triangulation to stay within the facial ROI by keeping only the tetrahedron inside the input mesh. Figure 3.5 depicts the DT constructed for a facial image, where red circles indicate the facial landmarks considered as vertices, and the blue lines represent the edges (The images are taken from CASMEII dataset with copyright permission©Xiaolan Fu). The DT helps provide optimum adjacent structures for AU matched points while also lowering the time costs of updating neighboring connections during frequent insertion and deletion operations. There are issues while indexing and matching with the DT, i.e., it may change the triangulations by introducing spurious triangles or deletion of essential triangles (Khodadoust and Khodadoust 2017). Thus, by employing AU indexes for the DT construction, the problem of DT is solved. Also, considering the triangles based on the facial indexes and its corresponding AUs reduces the memory and computational complexity.

For geometric feature extraction, initially euclidean distance (Cheddad et al. 2008) is computed between pair of triangular points as given in equation 3.1.

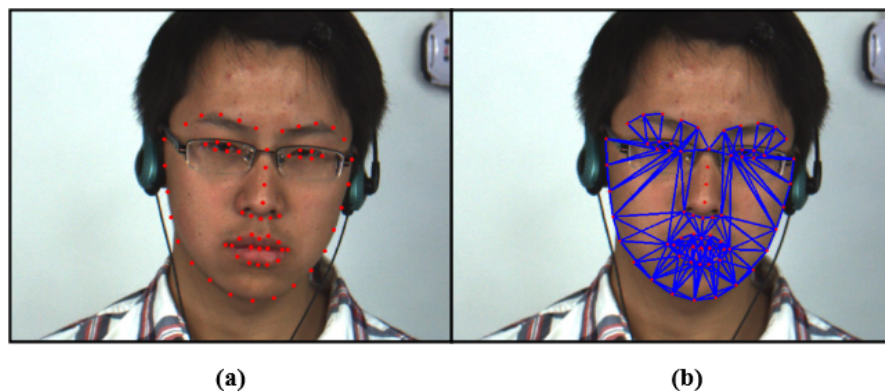


Figure 3.5: Facial landmarks points and Delaunay Triangulation (DT).

This study considers the area of the triangle computed using heron's formula given in equation 3.2, perimeter and circumradius calculated using equations 3.3 and 3.4 as a feature vector. The geometric features are computed for every AU, appended and stored into a CSV file. The feature-level difference between the onset and peak frames is calculated, which is used for further classification.

$$d(x, y) = \text{sqrt}((x_1 - x_2)^2 + (y_1 - y_2)^2) \quad (3.1)$$

where  $d(x,y)$  is the distance between two points  $x$  and  $y$ ; this calculation is done on three sides of the triangle namely  $a,b,c$ .

$$s = (a + b + c)/2.0 \quad (3.2)$$

where  $s$  is the perimeter and  $a,b,c$  are the sides of the triangle.

$$\text{area} = \text{sqrt}(s * (s - a) + (s - b) + (s - c)) \quad (3.3)$$

$$\text{circumradius} = (a * b * c)/(4.0 * \text{area}) \quad (3.4)$$

2. Facial Segmentation using Voronoi Diagram (VD) and Extraction of Texture Features: The approach is ultimately based on VD, and this technique is used in various fields and is useful in segmenting the facial region. VD is implemented to segment images based on facial landmark points. These points are carefully analyzed, identified, and mapped into distinct AUs (as per Table 3.2) based on the muscle movements. Thus, VD is based on extracting facial feature points, which is then used to define the centers of the Voronoi cells for the image segmentation as illustrated in Figure 3.6.

The first step is the initialization using landmark points (known as sites)



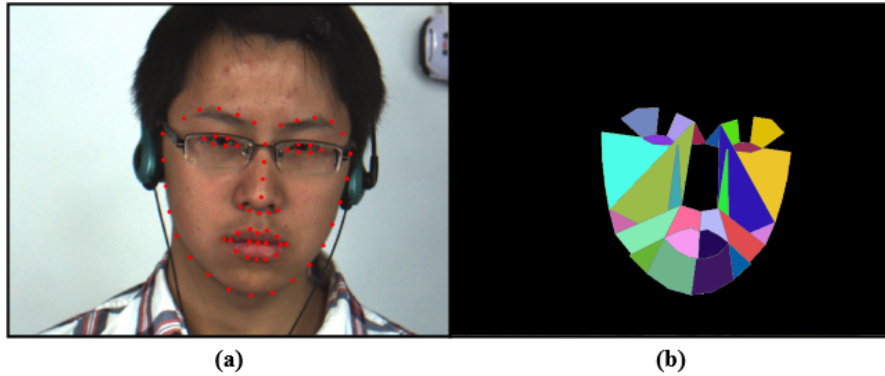


Figure 3.6: Facial landmark points and Voronoi Tessellation.

corresponding to AU regions. Then, generate the Voronoi Tessellations of the sites. The next step is the computation of clipped Voronoi Tessellation. The final step is the optimization, i.e., the position of the sites is updated until meeting user-defined criteria (minimize the Centroidal Voronoi Tessellation (CVT)). Once the Voronoi Tessellation is constructed for active AU regions, the texture features are extracted using the LBP (Ojala et al. 1996) approach. LBP histogram features are used as a texture feature because of its computational and simplicity reasons (Ghazouani 2021). This feature descriptor locates a keypoints within the image and generates the corresponding histogram distribution. The LBP operator outputs a binary code by calculating the difference between the central pixel and its equidistant circular neighbors as given in equation 3.5. The radius parameter  $r$  defines the distance, and the pixel parameter  $p$  denotes the number of neighbors (Ghazouani 2021).

$$LBP_{p,r} = \sum_{i=0}^{p-1} \mathbb{1}_R + (I(x_i, y_i) - I(x_c, y_c)) \cdot 2^i \quad (3.5)$$

Where,  $\mathbb{1}_R$  signifies the characteristic function of a subset  $R$ ,  $x_c$  and  $y_c$  are the coordinates of the central pixel,  $x_i$  and  $y_i$  are the coordinates of its  $i^{th}$  neighbor within the input image  $I$ . The basic  $LBP_{p,r}$  operator produces a  $2^p$  set of distinct output values that correspond to different binary patterns created by the  $p$  pixels. LBP features can capture micro-patterns, examine the micro-variations caused by wrinkles, and avoid the flaws of geometric features. So, appearance-based features are complementary to geometric distortions. Considering the combination of geometric and texture features helps design the more

discriminative features and thus helps in overcoming the FER challenges. The pixel-wise information is stored and then fed to the ensemble model for classification by finding the feature level difference between the onset and apex frames. The output generated during facial segmentation utilizing DT and VD approach based on AU indices is shown in Figures [3.7](#) and [3.8](#).

The facial landmarks are extracted and mapped into a particular AU, and this process is repeated for every individual AU. Once the segmented region is obtained, geometric and texture features are extracted and appended one by one for both onset and peak frames. Then both the features are ensembled, feature difference is calculated and fed into a distinct number of a machine learning model for classification.

A feature vector is created by combining the geometric and texture features retrieved from ROI complementing each other. Abundant subregions generated from DT and Voronoi Tessellations lead to a large number of local features. As a result, feature Selection is applied by selecting the AU regions which are not replicating. The AUs such as AU1, AU2, AU6, AU7, AU9, AU10, AU11, AU12, AU14, AU15, AU17, AU20, AU23, AU24, AU26 are examined based on the relevance of the regions (eyes, eyebrows, nose, mouth, chin, and cheeks) ([Wang et al. 2015](#)). Their respective geometric and texture properties (important features) are extracted, lowering the dimensionality. Before feature selection, a total of 11,412 features were obtained from all AU areas, with results that were similar in accuracy to those obtained with features created after feature selection. A total of 8160 features are developed after selecting non-overlapping AUs (regions) to avoid duplicating information. As a result, only a few critical AUs were needed to characterize MER efficiently.

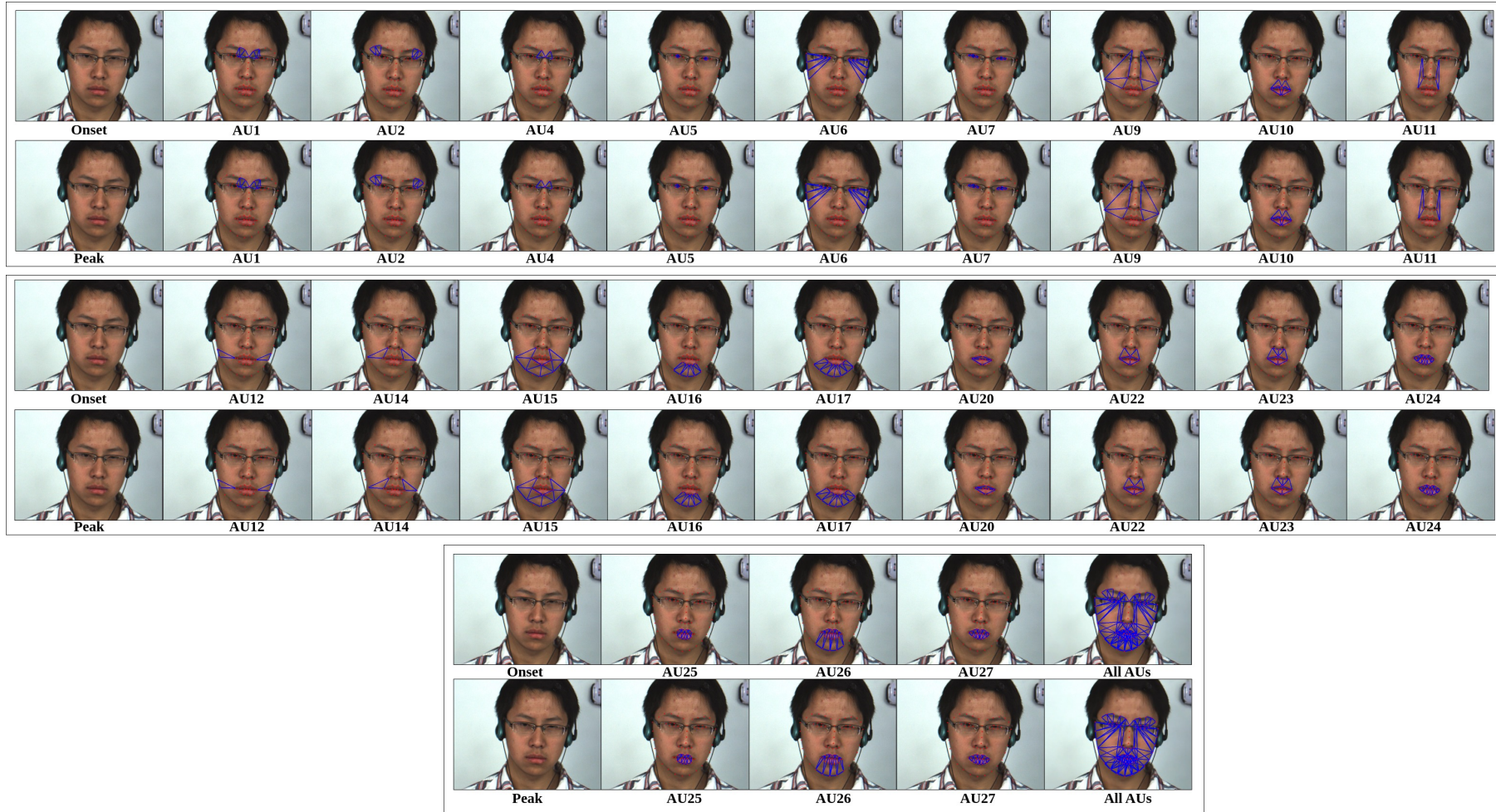


Figure 3.7: Delaunay Triangulation (DT) based facial segmentation using Action Unit (AU) indexes as the seed.

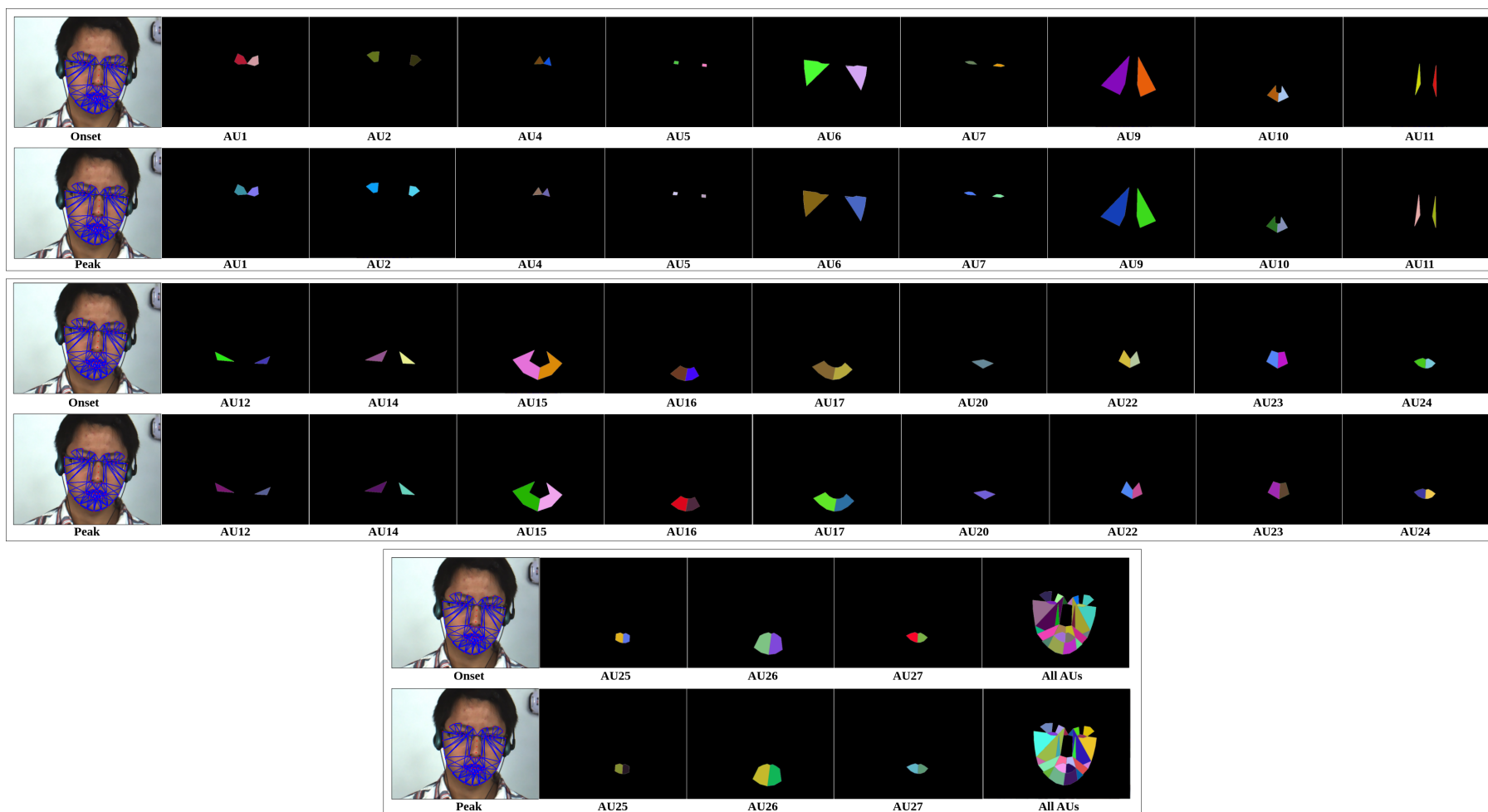


Figure 3.8: Voronoi based facial segmentation using Action Unit (AU) indexes as the seed.

### 3.3.3 Stage 3: Classification

For evaluating the proposed methodology, ensemble models like majority voting and stacking classifier methods (Aggarwal 2015; Malmasi and Dras 2018; Sakkis et al. 2001) are utilized to improve the classification performance of a single classifier. Various machine learning techniques are used for evaluation like Decision Trees (DTree) (Safavian and Landgrebe 1991), Extremely Randomized Trees (Extra Trees) (Yaddaden et al. 2018), Random Forest (RF) (Aggarwal 2015), KNN, Multi-Layer Perceptron (MLP), SVM (Dino and Abdulrazzaq 2019), Gaussian Naive Bayes (GNB) (McCallum et al. 1998). In the ensemble model, the experiments are carried out with various combinations of machine learning classifiers. The experiment uses stacking classifiers and a majority voting approach by combining three, five, and seven classifiers as ensemble models. In Figure 3.3 combination of three machine learning classifiers used for evaluation of the proposed model is represented as stacking classifier-I and majority voting-I; similarly, a combination of five and seven machine learning classifiers are represented as stacking classifier-II and majority voting-II, stacking classifier-III, majority voting-III, respectively. The reason behind choosing the various number of ensemble models is that each combination of machine learning gave different results when evaluated on three datasets.

- 1 Majority Voting Based Ensemble Classifier: In most FER cases, the facial image is associated with one emotion label. A majority of distribution label is obtained for an image using the ensemble classifier. In this experiment, the clubbed features are fed into the base classifiers, and machine-learning classifiers' predictions are aggregated. The majority voting strategy is utilized further to get the final prediction. The majority voting uses the predicted class labels with the highest votes to select a final class (Malmasi and Dras 2018).
- 2 Stacking Classifier: Stacking classifier combines multiple classifiers using meta learner (meta classifier) (Li and Zou 2017; Mihalcea 2002). This approach decreases the risk of getting varying output obtained from a distinct combination of classifiers. The base learner (base classifier (level 0)) takes the input and

makes predictions, i.e., the fused features are fed into level 0, and predictions are obtained. These predictions are fed into a meta classifier (level 1) with five cross-validations that analyze the pattern and predict the final output. The stacking classifier is considered as a powerful method compared to other ensemble models; it also reduces bias and variance (Aggarwal 2015).

## 3.4 RESULTS AND DISCUSSIONS

### 3.4.1 Database Description

Three databases are utilized for evaluation of the proposed model, Chinese Academy of Sciences Micro-Expression II (CASME II) (Yan et al. 2014), Spontaneous Actions and Micro-Movements (SAMM) (Davison et al. 2016) and Spontaneous Micro-expression Corpus (SMIC)-High speed (HS) dataset (Li et al. 2013; Pfister et al. 2011). The images of the SMIC-HS dataset are categorized into three classes negative, positive, and surprise. Due to the varying number of samples in ME categories, the expressions from CASME II and SAMM datasets are re-categorized into three expression classes: Happy is relabeled into positive class, disgusted and depressed classes are relabeled into negative class, surprise class remains unchanged, and other class is not considered for the evaluation.

### 3.4.2 Experimental Setup

The experiments were carried out using python language in an ubuntu 18.04 Long-term support (LTS) machine with a 2.60 gigahertz (GHZ) Central Processing Unit (CPU) and 64 GigaByte (GB) Random Access Memory (RAM). Offline data augmentations are performed on training samples. The test set contains the unseen subjects, which is different from that of the train set and has a high variability due to the differences in training and testing set and thus presents the meaningful differences in the estimation of the accuracy. In this experiment, CDE and HDE are performed on CASMEII, SAMM, and SMIC (HS) datasets. We do not claim to be the first to employ DT and the VD to recognize facial expressions. This work has utilized DT and VD to extract ROI based on AUs and improve the computational time. The proposed technique does not conduct facial cropping; instead, it uses the Dlib shape predictor to detect faces. However, if

there are any occlusions in the facial region, the shape predictor Dlib fails to recognize facial landmarks. The problem of facial occlusion needs to be solved, which will be done in future work.

The machine learning classifiers like Extra Trees, RF, DTree, SVM, KNN, MLP, and GNB are used for model classification and evaluation. These machine learning classifiers are examined independently to determine their performance by modifying the classifier's parameters (varying the depth of the tree, number of estimators, KNN value, and maximum iteration values). The best-performing classifiers were selected for the following stage as an ensemble classifier. The ensemble model utilized for the model's performance evaluation was the majority voting and stacking classifier.

The experiments are carried out using every individual classifier as a base and meta classifier with varying parameters. For example, RF is used as a base and meta classifier for evaluation with varying parameters. The results obtained when the same classifier is used as a base and meta classifier is presented in Table 3.3. The Cross-Database Evaluation (CDE) and Holdout Database Evaluation (HDE) result obtained when analyzing the proposed approach with the majority voting, and stacking classifier approach is presented in Figure 3.9. In this inquiry, the stacking classifier strategy to classification showed to be the most effective.

The performance metrics like accuracy, Unweighted Average Recall (UAR), and Unweighted F1 Score (UF1) are used to evaluate this experiment (Kumar et al. 2019). There is dataset imbalance with respect to the number of classes, thus UAR and UF1 score metrics is utilized for evaluation. The UF1 is a balanced F1, that is obtained by taking the average for each class F1-scores, and UAR is a balanced accuracy.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (3.6)$$

$$F1_c = \frac{2TP_c}{2TP_c + FP_c + FN_c} \quad (3.7)$$

$$UF1 = \frac{1}{N} \sum_{c=1}^N F1_c \quad (3.8)$$

### 3. Feature Extraction and Classification of Non-posed Expressions

$$UAR = \frac{1}{N} \sum_{c=1}^N \frac{TP_c}{M_c} \quad (3.9)$$

Where the terms like TP, TN, FP, and FN are true positive, true negative, false positive, and false negative; and  $TP_c$ ,  $FP_c$ ,  $FN_c$  are true positive, false positive, false negative for  $c^{th}$  class. The terms  $UF1_c$  is Unweighted F1 Score of  $c^{th}$  class; and  $M_c$  is the number of samples in  $c^{th}$  class.

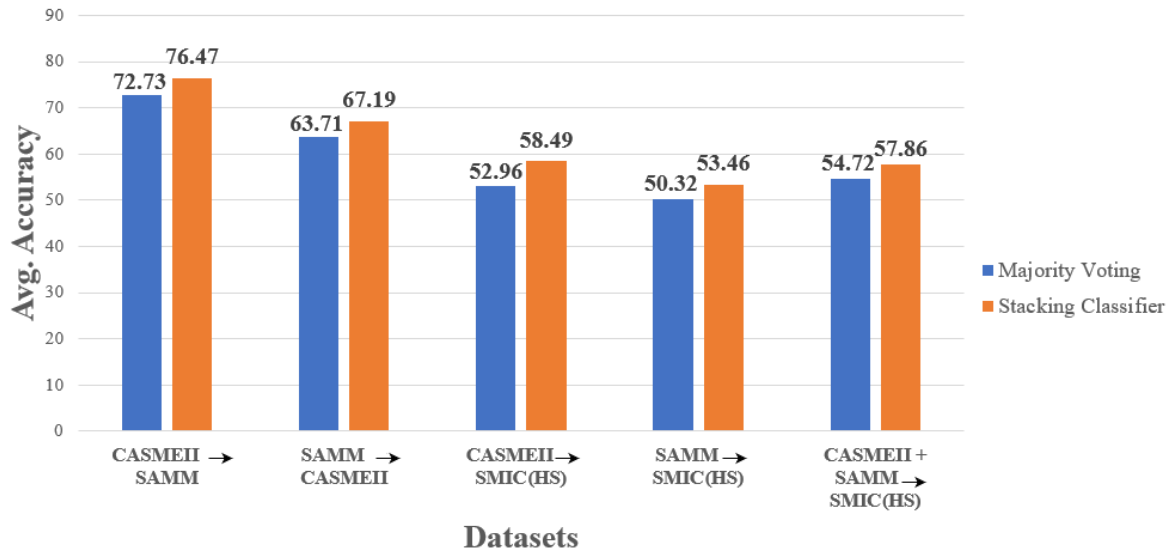


Figure 3.9: CDE and HDE result with majority voting and stacking classifier approaches.



Table 3.3: Comparison of results from different machine learning classifiers on three datasets.

*L. Algo.	*N	CASMEII→ SAMM		CASMEII→ SMIC (HS)		SAMM→ CASMEII		SAMM→ SMIC (HS)		CASMEII+ SAMM→ SMIC (HS)	
		Voting	Stacking Classifier	Voting	Stacking Classifier	Voting	Stacking Classifier	Voting	Stacking Classifier	Voting	Stacking Classifier
RF	3	72.22%	75.53%	50.75%	51.12%	59.78%	63.10%	45.29%	51.06%	49.77%	51.11%
	5	72.27%	72.10%	50.79%	51.54%	59.88%	62.86%	46.88%	51.33%	50.48%	52.13%
	7	72.48%	71.97%	50.83%	51.83%	59.83%	63.15%	46.05%	51.16%	50.28%	52.05%
DTree	3	72.16%	76.16%	52.96%	56.51%	58.07%	62.56%	47.38%	52.12%	51.12%	51.22%
	5	70.43%	<b>76.47%</b>	52.91%	<b>58.49%</b>	59.09%	62.32%	47.15%	51.51%	50.79%	52.43%
	7	70.27%	76.38%	52.76%	57.56%	58.99%	62.66%	47.27%	51.67%	50.72%	53.15%
Extra Trees	3	72.68%	74.04%	50.06%	51.12%	59.71%	63.35%	49.12%	51.16%	47.12%	51.12%
	5	72.73%	75.53%	50.11%	51.53%	60.95%	63.44%	49.45%	51.58%	48.25%	51.54%
	7	72.52%	74.72%	50.53%	51.26%	59.61%	63.35%	48.96%	51.44%	48.36%	52.75%
SVM	3	70.01%	75.63%	50.56%	53.51%	63.10%	66.15%	50.01%	52.95%	52.84%	55.79%
	5	71.70%	75.32%	50.56%	53.97%	63.71%	<b>67.19%</b>	50.32%	52.79%	53.02%	<b>57.86%</b>
	7	70.20%	75.27%	50.12%	52.79%	63.32%	66.95%	50.11%	<b>53.46%</b>	54.72%	57.37%
MLP	3	65.19%	74.53%	48.63%	50.76%	59.39%	62.41%	48.39%	50.64%	47.22%	52.36%
	5	66.52%	75.34%	47.63%	49.39%	59.95%	62.65%	47.60%	49.06%	48.68%	53.88%
	7	65.06%	74.02%	47.26%	49.57%	59.24%	61.39%	48.35%	50.05%	48.95%	54.02%
KNN	3	57.91%	68.53%	47.31%	48.32%	52.29%	58.64%	46.38%	51.95%	47.32%	49.75%
	5	57.18%	69.16%	46.80%	48.78%	57.95%	59.28%	45.39%	51.78%	47.82%	48.58%
	7	57.96%	68.21%	47.21%	48.62%	55.74%	58.73%	46.73%	51.04%	47.44%	48.82%
GNB	3	58.13%	68.26%	42.49%	51.91%	52.79%	58.56%	44.64%	50.41%	45.66%	48.92%
	5	58.65%	68.62%	43.93%	50.92%	52.16%	58.72%	45.79%	50.54%	45.72%	49.06%
	7	57.23%	68.16%	43.61%	51.23%	52.94%	57.93%	45.72%	50.57%	46.37%	49.91%

\*L. Algo.-Learning Algorithms, \*N-Number of Ensemble Classifiers

### 3.4.3 Comparison with state-of-the-art methods

The proposed approach is compared with other state-of-the-art methods and presented in Table 3.4. The confusion matrix results obtained with CASMEII, SMM, and

Table 3.4: Comparison of MER performance against state-of-the-art methods evaluated on CASMEII, SMM, SMIC(HS) datasets.

Reference	Source Dataset	Target Dataset	Model	*Exp	*Acc	UAR	UF1
Zong et al. (2018)	CASMEII	SMIC(HS)	SVM	3	*NA	0.3791	NA
	CASMEII	SMIC(HS)	DRLS		NA	0.5465	NA
Kumar et al. (2019)	CASMEII	SMM	CNN, EMM	3	NA	0.6506	0.6595
	SMM	CASMEII			NA	0.6044	0.6210
	CASMEII	SMIC(HS)			NA	0.6684	0.6770
	SMM	SMIC(HS)			NA	0.5846	0.5924
Peng et al. (2019)	CASMEII + SMM	SMIC(HS)	ATNet	3	NA	0.503	0.524
Xie et al. (2020)	CASMEII	SMIC(HS)	AU-GACN	3	34.4%	NA	NA
	SMM	SMIC(HS)			45.1%	NA	NA
Zhang et al. (2021)	CASMEII	SMIC(HS)	IDSDA	3	57.93%	NA	NA
Takalkar et al. (2021)	CASMEII	SMM	LGAttNet	NA	73.2%	NA	NA
	SMM	CASMEII			67%	NA	NA
Yanliang et al. (2021)	CASMEII	SMIC(HS)	SLJDA	3	NA	0.5546	NA
Proposed Methodology After Feature Selection	CASMEII	SMM	Ensemble Model	3	76.47%	0.6872	0.6851
	SMM	CASMEII			67.19%	0.6454	0.6449
	CASMEII	SMIC(HS)			58.49%	0.5728	0.5754
	SMM	SMIC(HS)			53.46%	0.5225	0.5239
	CASMEII + SMM	SMIC(HS)			57.86%	0.5639	0.5676

\*NA–Not Available; \*Exp–Expressions; \*Acc–Accuracy

SMIC datasets are presented in Figure 3.10. The results obtained suggest that the proposed model achieves promising results on three ME datasets. Due to the data size constraint, deep learning methods performed poorly. By employing a DT and VD approach to segment the face based on AU points, extracting geometric and texture features, and feeding to the machine learning model (stacking classifier model) suited best to adapt to unknown test data. The extraction of relevant features from the important facial region based on AU indexes greatly enhanced the ME-related features and aided the recognition of MEs.

Comparing the results presented in Table 3.4, it is observed that when utilizing

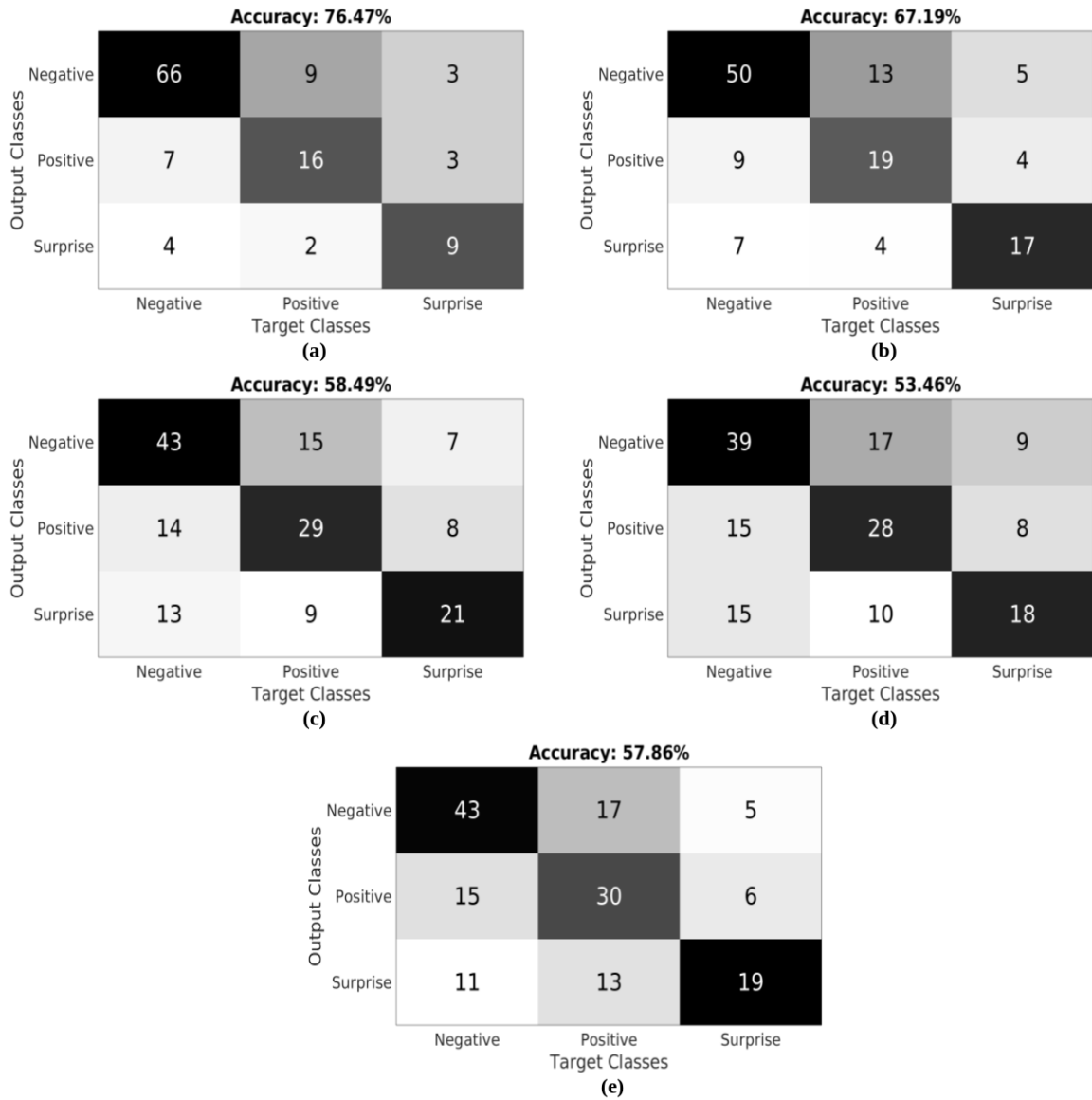


Figure 3.10: Confusion Matrices (a) CASMEII→SAMM (b) SAMM→CASMEII (c) CASMEII→SMIC (d) SAMM→SMIC (e) CASMEII+SAMM→SMIC

CASMEII as a training set and the SAMM as a test set, accuracy improved by 3.27% compared to [Takalkar et al. \(2021\)](#). In addition when compared to [Kumar et al. \(2019\)](#) UAR and UF1 scores improved by 0.0366 and 0.0256, respectively. Similarly, when SAMM is used as a training set and CASMEII as a test set, the accuracy increased by 0.19% compared to [Takalkar et al. \(2021\)](#). The UAR and UF1 scores also increased by 0.041 and 0.0239 respectively compared to [Kumar et al. \(2019\)](#).

Using CASMEII as training set and SMIC as a test set the accuracy increased by 0.56% compared to [Zhang et al. \(2021\)](#) and also UAR score increased by 0.0182

compared to Yanliang et al. (2021). But, the performance of UAR and UF1 scores is decreased by 0.0956 and 0.1016 respectively compared to Kumar et al. (2019). Similarly, when SAMM is utilized as a training set and SMIC as a test set, the performance of UAR and UF1 scores is reduced by 0.0621 and 0.0685 compared to Kumar et al. (2019). The experiment also employs HDE, a stricter cross-validation procedure where CASMEII and SAMM datasets are utilized as the training set and SMIC as the test set. As a result, UAR and UF1 scores are improved by 0.0609 and 0.0436, respectively, as compared to Peng et al. (2019).

### 3.5 SUMMARY

In this chapter, a novel MER system is proposed that utilizes DT and VD approach based on the AUs description. Distinct AUs are modeled, and their corresponding geometric and texture features are extracted and integrated, which help in recognition of minute changes over the face. The extracted features are appended and fed into an ensemble of ML classifiers for further classification based on majority voting and stacking classifier approaches. The experiment results gave good results on non-posed datasets.

In this chapter, a novel ME method that uses discriminative characteristics extracted from AU regions to help in ME identification, overcoming challenges like subtle variations in a short time and intricate interplay between facial areas. The proposed methodology incorporates the DT and VD approach for facial segmentation and extracting geometric and texture features from those regions. The DT and VD approach to retrieve ROIs, geometric, and texture feature ensembles complemented each other. The performance showed efficient results in distinguishing minute changes in facial areas and ME. Finally, the ensemble of machine learning classifiers proved to be efficient in classifying ME data. To our knowledge, this is the first study to employ AU indexes as a seed for DT and VD to segment the face and extract essential features for MER. Finally, experimental results on three ME datasets gave good results.

The CDE conducted on three databases, CASMEII, SAMM, and SMIC(HS), proved the robustness of the proposed model and thus showed the model can be helpful for real-

time processing. The accuracy of 76.47% is obtained when evaluated using CASMEII as a training set and SAMM as a test set. When the SAMM dataset is utilized as a training set and CASEMII as a test set, improvement by 0.19% accuracy is obtained compared to [Takalkar et al. \(2021\)](#). Also, the HDE evaluation attained UAR and UF1 scores by 6.09% and 4.36%, proving the robustness of the model. Future work will focus on more complex algorithms for detecting facial landmark points, extracting more relevant data, and making it practical to use in real-world applications. Improve the performance on SMIC datasets, extend the work on video sequences, and develop a better feature extraction strategy for MER in live stream videos. Chapter 4 gives the implementation details of facial occlusion detection that is required for the recognition of facial region and expressions efficiently.



## CHAPTER 4

### FACIAL OCCLUSION DETECTION

This chapter briefly overviews the localization mechanism utilized to detect occluded facial regions.

#### 4.1 INTRODUCTION

Rapid advancements in deep learning have produced some of the most favorable outcomes for face recognition systems. However, they perform inadequately under unconstrained environments when facing occlusions, varying facial expressions and poses, inadequate resolution, and lighting conditions. A common belief is that facial occlusions are among the most challenging issues to solve. The occlusion mainly covers a significant portion and eradicates efficient features of the face (Hemathilaka and Aponso 2022). The occlusion blocks the facial region, making it more challenging to extract distinguishing features. Occlusions frequently occur in natural settings and are challenging in computer vision and object detection. Since the occluded area of a face image might vary in position, size, and form, facial identification presents a challenging problem. The problem of handling facial occlusions is broadly classified as a holistic approach and a part-based approach (Lahasan et al. 2019). In the holistic approach, the image of the face is treated as a whole entity, and global information is utilized to carry out face recognition. In contrast, the facial image is divided into overlapping or non-overlapping segments in a part-based approach. Each segment is handled separately, and local regions are used

for matching rather than considering the whole face entity.

According to [Zeng et al. \(2021\)](#), occluded face images range from real occlusions to the least real (synthetic) images; as a result, there are five testing scenarios for occluded face identification as depicted in Figure [4.1](#). Faces in actual occlusion have realistic features like a scarf, mask, sunglasses, hair, and other accessories. Images of faces that are only partially visible, possibly lacking some facial features, are known as partial faces. Images created to look like real occlusions are seen in synthetic occlusions ([Jiang et al. 2022](#)). Occluding Rectangle will block the faces with white and black rectangles to produce an obscured face. When obscuring unrelated photographs with bottles, veggies, fruits, etc., the facial images are covered up with irrelevant images.

Three types of face recognition can be performed when occlusions are present ([Zeng et al. 2021](#)) as depicted in Figure [4.1](#). The first is Occlusion Robust Feature Extraction (ORFE), which reduces the complexity of identifying the occluded face by extracting features (patch and learning-based features) that are resistant to occlusion and less impacted by outliers. Occlusion-Aware Face Recognition (OAFR) falls under the second category. The first set of methods in this category detects the occlusion before deriving a representation from the unobscured regions ([Jiang et al. 2022](#)). The second method is partial face recognition, where occlusion detection is not taken into account, and just a partial face is available to identify faces. Occlusion Recovery-based Facial Recognition (ORecFR) is the third type of face recognition; occlusion recovery is regarded as a substitution for solving the occlusion issue in the image space. The occluded image is repaired using a reconstruction or inpainting approach ([Nazeri et al. 2019](#)), and the face recognition task is performed.

Facial occlusion is one of the most complex problems, as there needs to be prior knowledge of the occluded area, where it might appear, and what size or form it may take in a facial image. Gathering a large-scale training dataset containing every form and distinct type of occlusions in a realistic scenario is not viable. Thus, face recognition under occlusions remains an open research topic to this day ([Zeng et al. 2021](#)). Different scenarios exist, like facial accessories, self-occlusions, external



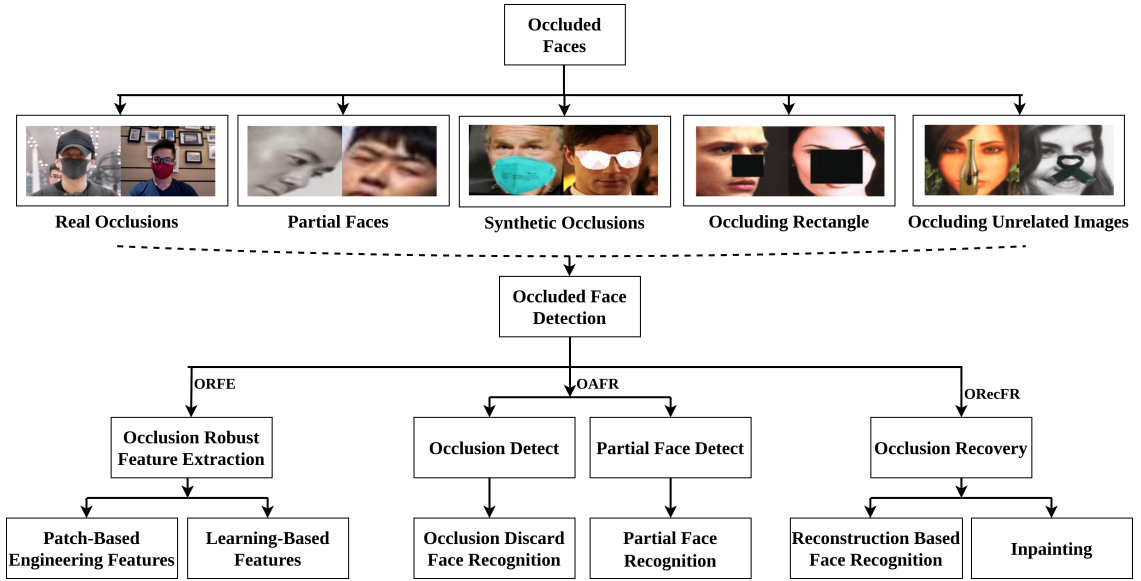


Figure 4.1: Categories of face recognition during the presence of facial occlusions.

occlusions, partial occlusions, artificial occlusions, and intense illumination variation (Zeng et al. 2021).

The current state-of-the-art methods still fall short with severe occlusions, despite advancements with normal or weakly obscured faces (Huang et al. 2022). Occlusions suppress the face’s natural landmarks, lowering facial recognition’s effectiveness. The detection of facial occlusion is the central goal of this work. Compared to engineered features, learned features are more adaptable when occlusions exist in diverse locations (Zeng et al. 2021).

Face occlusion detection is an essential task to avoid crimes, and this problem includes successive steps such as localizing the face, segmentation, extraction of vital features, and recognition (Xia et al. 2015). This work proposes a robust localization approach using the Xception network with a residual attention (Xcep-RA) mechanism to detect facial occlusions effectively. The occlusions detected by the proposed model are visualized using the Gradient weighted Class Activation Map (Grad-CAM) method. In contrast to previous works, this work focuses on detecting facial occlusions using the Xcep-RA mechanism to further aid the effectiveness of feature selection, leading to more accurate recognition.

The contributions of the chapter are :

- The proposed methodology utilized a residual attention mechanism within the Xception network (Xcep-RA) to effectively detect facial occlusions.
- The Grad-CAM visualization technique is utilized along with the proposed approach to visualize the detection of facial occlusions.

The chapter is organized as follows: Section 4.1 is for introduction, Section 4.2 briefs the overview of Xception Network architecture, Section 4.3 describes the proposed hybrid CNN model and its block diagram, Section 4.4 presents the results and its detailed analysis using various evaluation parameters, and Section 4.5 summarizes the proposed method and its significance.

## 4.2 PRELIMINARIES

CNN is a deep learning framework for image classification. It includes convolution and pooling layer operations that extract high-dimensional characteristics from the images. While pooling operations are employed to decrease the number of features by enhancing the model's resilience. The Xception model is a CNN (Dodia et al. 2022). Xception network is an extension of inception architecture and was proposed by Francois Chollet (Chollet 2017). The depthwise separable convolutional layer (Venkatesh and Koolagudi 2022) and residual structure is included within the Xception network to gain superior performance in recognizing facial occlusions (Chollet 2017; Liu et al. 2022). The term Xception refers to "Extreme Inception". The Extreme Inception module and a depthwise separable convolution module, which maps cross-channel correlations using  $1 \times 1$  convolution, are nearly identical. In the Inception module, operations like  $1 \times 1$  convolution are carried out first, whereas, in depthwise separable convolution, channel-wise spatial convolution is carried out first, followed by  $1 \times 1$  convolutions (Chollet 2017).

The Xception architecture includes 36 convolutional layers structured into 14 modules (entry, middle, and exit flows) utilized for feature extraction, followed by a fully connected layer; all modules except the first and last comprise linear residual

connections. Data processing occurs in the entry, middle, and exit flows. The main core of this network is depthwise separable convolution (Le et al. 2021); this layer diminishes the network complexity and maximizes information transfer between the layers (Liu et al. 2022). The linear combination of the depthwise separable convolution with residual connections makes the model structure easier to modify (Chollet 2017). The underlying hypothesis of the Xception architecture is strong to decouple the spatial convolutions and cross-channel correlations. The input data is translated into spatial correlations individually at each output channel, and  $1 \times 1$  depthwise correlation is performed that captures the cross-channel correlation (Dodia et al. 2022). Thus, it provides slightly better performance. In the proposed architecture, three residual attention modules are added between these 14 modules of the Xception network. The Xcep-RA mechanism increases the network's capacity to extract global information.

### 4.3 PROPOSED MODEL FOR FACIAL OCCLUSION DETECTION

The first step for detecting occlusions is localizing occluded objects; the second step is to either remove occlusions or restore the occluded region (Dagnes et al. 2018). This work focuses on detecting a subset of typical occlusions, especially those caused by external accessories such as facial or surgical masks and pairs of glasses.

The proposed model adapts the Xception network, which consists of 14 modules made up of 36 convolutional layers that are all connected linearly around them except for the first and end modules. Blocks in each flow comprise the Batch Normalization, ReLU activation function, 33 depthwise separable convolutional kernels, and the max pooling layer. At the same time, Batch Normalization and 11 convolutional kernels comprise the residual structure of entry and exit flow. The residual attention mechanism used within the architecture further improves the network model's performance by eliminating false gradients to update network parameters, increasing meaningful features, and impeding meaningless information. The overall pipeline of the architecture is presented in Figure 4.2 and explained briefly further.

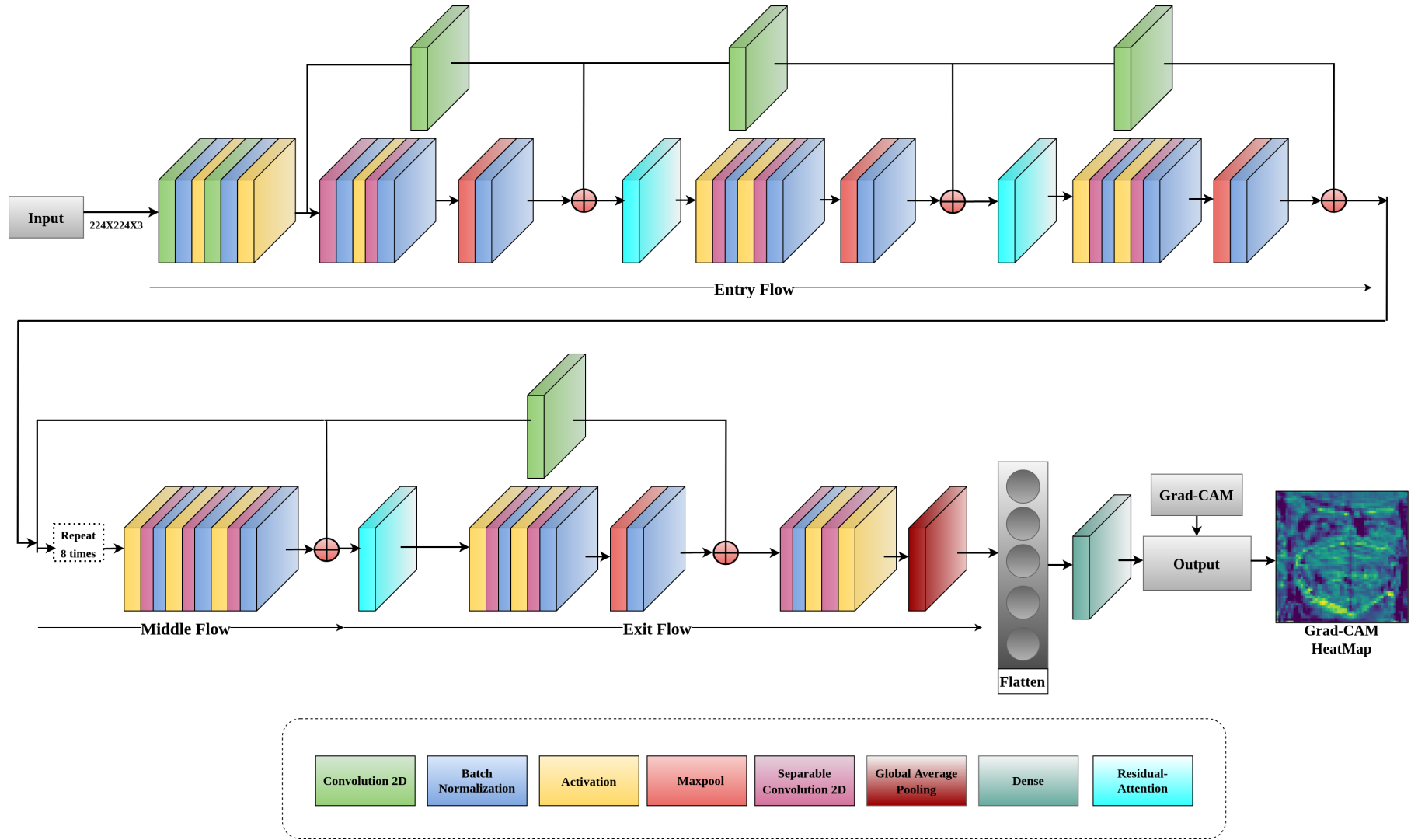


Figure 4.2: The overall network structure for the localization of facial occlusions.

The input data is pre-processed into  $224 \times 224$  resolution size, which is fed into the proposed Xception module. In the proposed approach, the Xception network is modified with a residual attention mechanism to detect facial occlusions efficiently. The residual attention mechanism optimizes the network model to prevent erroneous gradients from being utilized to update the network parameters (Liu et al. 2022). As a result, the meaningful features are enhanced while the ineffective data is suppressed. The residual attention module is added behind the  $2^{nd}$ ,  $3^{rd}$ , and  $12^{th}$  layers of the Xception model (Liu et al. 2022). The Trunk Branch (TB) and the Soft Mask Branch (SMB) are the two branches that make up the residual attention module (Wang et al. 2017). The TB branch is primarily employed for feature processing, while the SMB feature selector serves as a filter during gradient updates. The attention module's stability is improved, and the impact of noise on gradient updates is reduced using the SMB branch.

The entry flow of the Xception module contains four blocks, the middle flow has eight blocks, and the exit flow comprises two blocks. Adding residual attention blocks according to the change in feature map size enhances the performance by extracting good features and suppressing irrelevant data (Liu et al. 2022). Inspired by Liu et al. (2022), optimal results are obtained by the proposed model adding residual attention blocks behind  $2^{nd}$ ,  $3^{rd}$ , and  $12^{th}$  layer. Further, the proposed work utilizes Grad-CAM (Selvaraju et al. 2017) visualization technique to detect and localize the presence of the occlusion. The heatmap is generated using Grad-CAM and the proposed Xception-based classifier. The Grad-CAM (Selvaraju et al. 2017) is a high-class discriminative approach with high resolution. The final convolutional layer often keeps the spatial information typically lost in the fully-connected layer. The final convolution layer presents high-level semantics (class-specific knowledge) and precise spatial data. Grad-CAM then assigns significance values to each neuron for a specific decision of interest, capturing the gradient information coming to the final convolutional layer. The proposed work performs simultaneous categorization of an occluded and non-occluded task, as well as occlusion detection. The primary goal of this study is to identify the occluded faces efficiently.

### 4.4 RESULTS AND DISCUSSIONS

#### 4.4.1 Dataset Description

The experiments are conducted using three datasets, Webface-OCC (Huang et al. 2021), LFW (Learned-Miller et al. 2016) and RMFD (Wang et al. 2023). In this work, 5000 images from the unmasked and masked classes are used to balance the images in the two categories of RMFD datasets. The images of the input are resized to  $224 \times 224$ . The Webface-OCC database is utilized for training the Xcep-RA mechanism.

#### 4.4.2 Experiment Analysis and Comparisons

Python programming language is used for implementing the proposed system, using packages such as Keras and Tensorflow (Gulli and Pal 2017). The proposed model was executed up to 50 epochs. The adam optimizer was used in the training phase to update the weights. A batch size of 32 was used. To reduce the computations and to address the vanishing gradient problem, ReLU (Nair and Hinton 2010) was utilized as an activation function within the proposed architecture. The categorical cross-entropy is utilized as a loss function of the model. At the final layer, the sigmoid function is applied for classification. The cross-dataset evaluation is used in the proposed study to address the issue of the lack of occluded face dataset (Chen et al. 2018). LFW and RMFD datasets are used for testing, and the Webface-OCC dataset is used for training. For the evaluation of the Webface-OCC dataset, an 80/20 split between the training and test sets is taken into account.

The confusion matrix obtained for WebFace-OCC, LFW, and RMFD datasets is depicted in Figure 4.3. The performance metrics like accuracy, specificity, sensitivity, F1-score, and precision are utilized to evaluate the Xcep-RA mechanism. The results obtained from the confusion matrices are presented in Table 4.1. Comparing Xcep-RA with the Xception network without residual block, there is an improvement in the performance of the results. Thus, it can be observed that incorporating a residual attention mechanism into the Xception network improves the model's detection rate even further. The Xcep-RA module for detecting facial occlusion produced accuracies

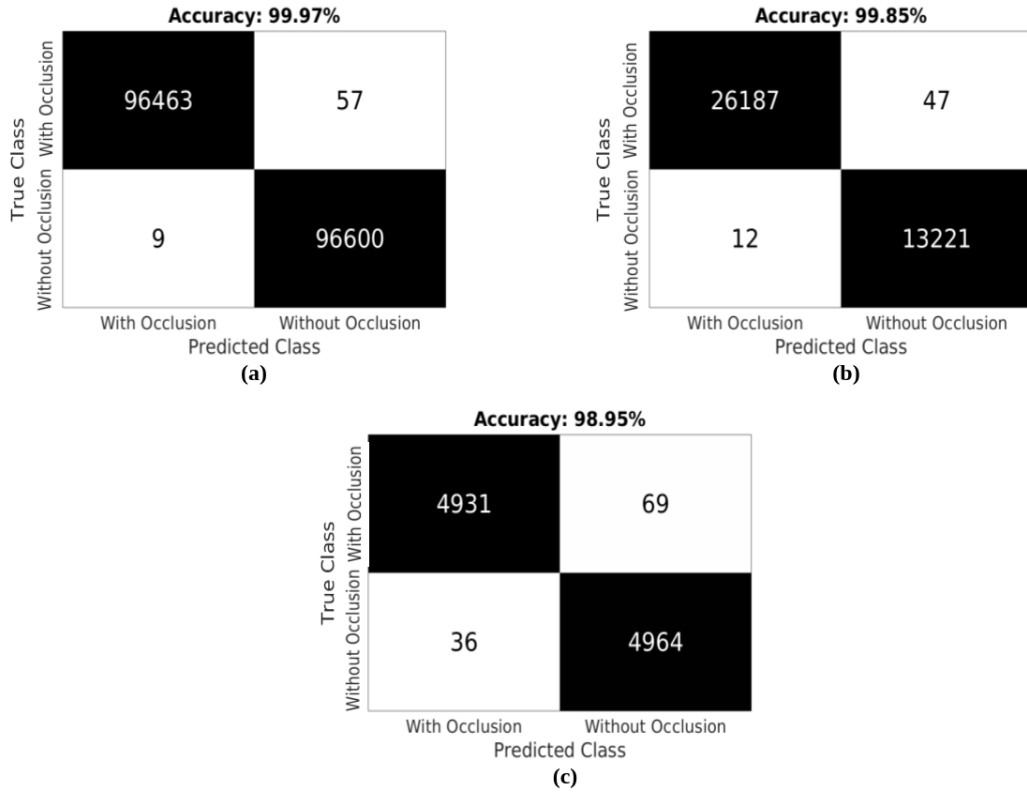


Figure 4.3: Confusion matrix results for the datasets (a) Webface-OCC (b) LFW (c) RMFD

Table 4.1: Experiment results obtained with the proposed methodologies on three datasets.

Dataset	Method	Accuracy	Specificity	Sensitivity	F1-score	Precision
Webface-OCC	Xception Model	96.54%	96.19%	96.90%	96.53%	96.15%
LFW-mask		96.05%	92.63%	97.87%	97%	96.14%
RMFD		97.21%	96.37%	98.09%	97.16%	96.26%
Webface-OCC	Xception with residual attention mechanism (Xcep-RA)	99.97%	99.99%	99.94%	99.97%	99.99%
LFW-mask		99.85%	99.91%	99.82%	99.89%	99.95%
RMFD		98.95%	99.28%	98.62%	98.95%	99.28%

of 99.97%, 99.85%, and 98.95%; the specificity of 99.99%, 99.91%, and 99.28%; sensitivity of 99.94%, 99.82%, and 98.62%; F1-score of 99.97%, 99.89%, and 98.95%; the precision of 99.99%, 99.95%, and 99.28% on Webface-OCC, LFW and,

RMFD datasets respectively.

According to studies and analysis, the mouth and the eyes are the two face surfaces that are the most discriminating and help us recognize facial expressions (Abate et al. 2023). The COVID-19 pandemic constraints have also shown that state-of-the-art methods for analyzing the face might suffer fatal failures because of the occlusions of utilizing facial masks. Furthermore, occluded eyes also cause a significant drop in performance, though it is not as severe as when facial masks are present. Thus, it confirms that, just like with face biometric recognition, occluded faces by facial masks continue to be a difficult obstacle for computer vision solutions.

##### 4.4.2.1 Occlusion Detection

The effect of utilizing the Xcep-RA to localize the facial occlusions is depicted in Figure 4.4. The facial occlusions are detected, and results are displayed layerwise. The module 3 heatmap visualization is displayed first; as the layer advances, face occlusions are better localized for detection. Occluded sections are visible in the heatmap created in module 14.



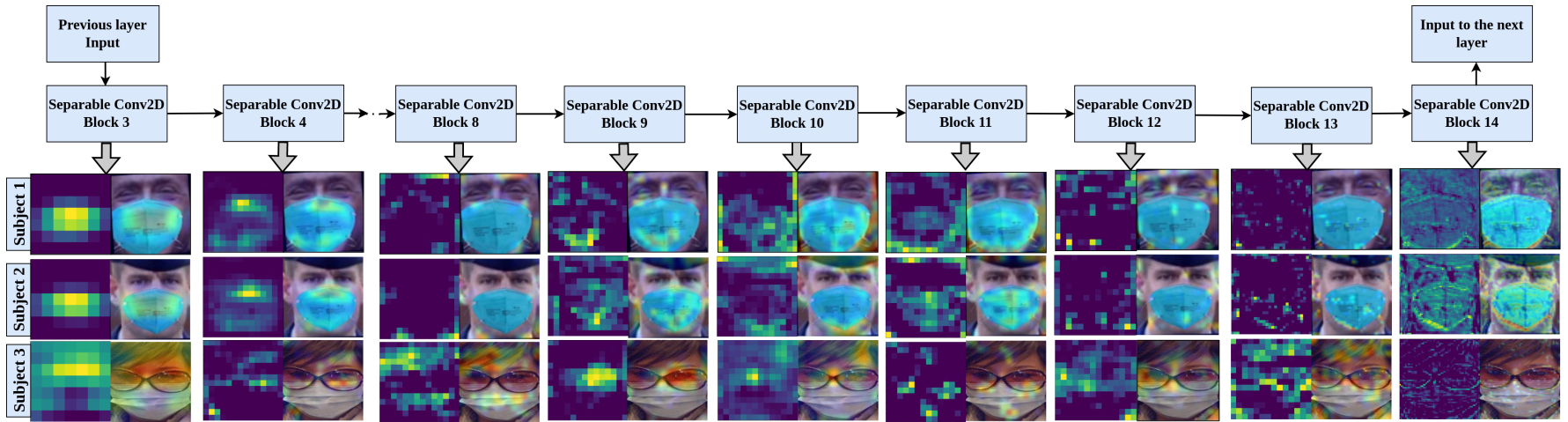


Figure 4.4: The localization of facial occlusions using Grad-CAM visualization approach.

#### 4.4.2.2 Comparison with State-of-the-art Methods

This work utilizes an Xcep-RA unit to detect facial occlusions. The key benefit of using the Xcep-RA module is the model detects and localizes the facial occlusions effectively. The Webface-OCC dataset is utilized as a training set. The performance of the proposed model is compared with previous research works and presented in Table 4.2. Compared to Huang et al. (2021) using the LFW-mask and RMFD datasets, the proposed model performs 2.77% and 20.7% better, respectively. Additionally, the model outperforms Song et al. (2019), and Wan and Chen (2017) on the LFW-mask dataset by 0.65% and 3.55%, respectively. The RMFRD and LFW-mask datasets show an increase of 12.55% and 5.27%, respectively, compared to Wang et al. (2023), and the LFW-mask and RMFD datasets show a decrease of 0.15 and 0.69 percent compared to Loey et al. (2021), respectively. The degradation in the performance may be due to the variation in the image sample considered for each class. The performance may be improved by applying image processing techniques to enhance the image's quality, remove background noise, etc.

Table 4.2: Comparison of the proposed model with previous approaches.

Dataset	Methods used	Accuracy
LFW-mask (Huang et al. 2021)	ArcFace method for face recognition	97.08%
RMFD (Huang et al. 2021)		78.25%
LFW-mask (Loey et al. 2021)	ResNet-50 & classical machine learning models	100%
RMFD (Loey et al. 2021)		99.64%
LFW Benchmark (Song et al. 2019)	Pairwise Differential Siamese Network	99.20%
LFW-mask (Wan and Chen 2017)	Mask-ResNet	96.3%
RMFRD (Wang et al. 2023)	Arcface (ResNet 50 is a backbone network)	86.40%
LFW-mask (Wang et al. 2023)		94.58%
LFW-mask	<b>Proposed Model</b>	<b>99.85%</b>
RMFD		<b>98.95%</b>

## **4.5 SUMMARY**

Facial occlusions occur due to obstacles covering the frontal face increasing the difficulty of the model in extracting the discriminative features. The occlusions block the facial region, thus leading to registration errors and inaccurate face alignment. The primary goal of this research is to identify facial occlusions and minimize data loss during face recognition. The Xcep-RA and Grad-CAM visualization approach is used in this study to find occlusions and combat erroneous predictions.

Three datasets are utilized to detect the facial occlusions and evaluate the proposed model. A comparative result has been carried out with previous works. The proposed model achieved an increased performance on LFW-mask and RMFD datasets compared to a few associated works in testing accuracy. Future work would be to enhance the performance and extend this work to identify faces and facial expressions accurately. Also, developing a unified framework combining occlusion detection and face recognition from occluded faces and focusing on an automatic face recognition system. Chapter 5 discusses on the approach proposed for the recognition of posed expressions.



## CHAPTER 5

### CLASSIFICATION OF POSED EXPRESSIONS

The identification of posed expressions is briefly covered in this chapter. In the experiment, intermediate features are retrieved, and ensemble techniques are used to classify the posed expressions.

#### 5.1 INTRODUCTION

Every region of the face conveys some important affective information. Sometimes, it is challenging to separate the same subject's facial features in two different expressions, as they may share the same feature space (Lopes et al. 2017). There are issues with selecting appropriate features to distinguish individuals' emotions from various categories of emotions (Zhong et al. 2014). Expressions keep varying within the same culture (Dailey et al. 2010; Russell 1991), and patterns may depend on environment settings, mood, and situations, making it difficult for machines to recognize them efficiently (Lopes et al. 2017). Variations in the face, facial occlusions, head poses, and illumination also degrades the overall system's performance. A generalized approach is needed that could overcome all these variations and help in building an efficient, robust system for recognition of expressions (Feifei et al. 2018).

The composition of both feature extraction and classification techniques is essential for FER. The key challenges in efficiently recognizing facial expressions are a selection of efficient feature extraction and classification techniques. If features are

sparse, the best classifier's performance would also gradually decrease (Wen et al. 2017). Handcrafted features like SIFT, Gabor, LBP, and HOG have achieved a breakthrough in various fields. These handcrafted low-level features work well on a small amount of training data and are inadequate for extracting discriminative information. It is arduous to fine-tune these low-level features according to input data. These disadvantages of low-level features made it inefficient in recognizing facial expressions accurately in real-world applications. Thus, the deep learning models overcame these challenges and automatically learn from raw data, represent the data on multiple levels, and contain more abstract information. The rapid advancement in the deep learning field has impacted FER and has shown promising results in recognition of facial expressions.

This work explores the EfficientNet (Tan and Le 2019) model, the high-quality model from the CNN group of models, to efficiently recognize facial expression from static images. It is efficient in terms of a lesser parameter (4M parameters) and achieves an increased performance of 2.29% to 10.71% with multi-pose and frontal pose data compared to previous CNN models. The EfficientNet B0 baseline architecture is used as a feature extractor to improve the FER system's recognition rate and accuracy. Further, the features extracted from the EfficientNets intermediate layer are fed to machine learning classifiers for classification. A combination of deep learning models and machine learning classifiers effectively improves the ability of classification algorithms. Two ensemble models, EfficientNet B0, features fed to stacking classifier (EfficientNet Model with the Stacking Classifier (SC-EffNet)), and EfficientNet B0 features provided to machine learning classifiers based on the frequency of votes (EfficientNet Model with the Frequency-based Voting strategy (FV-EffNet)), are proposed to classify facial expressions into respective expression classes. Thus, the combination of multiple classifiers induces higher-level classifiers and tries to learn all possible patterns from the base classifiers (Sakkis et al. 2001), which further enhances the overall performance of FER.

The proposed work uses the EfficientNet model, which is computationally and memory-efficient compared to previous CNN models. The intermediate features of the

models fed to machine learning classifiers using a frequency-based approach improved the system's accuracy even further. As a result, we chose the top five best weights and their corresponding intermediate features, which we fed into machine learning classifiers to test the system's performance. Additionally, to improve the model's performance, a stacking classifier (an ensemble approach) was used. The meta classifier analyzed the pattern of base classifiers and learned from their errors before making the final prediction. By integrating the outputs of base classifiers, the ensemble model makes accurate predictions, reduces over-fitting, reduces the risk of selecting a single classifier, and achieves good results. There is no work related to FER using best model weights, the extraction of EfficientNet features, and the stacking classifier for classification to the best of our knowledge.

In affective computing, emotion recognition from video data is the current issue (Renda et al. 2019). Even though the amount of information obtained from the video signals is comparatively more. The quickness and variability in dynamics (rapid changes in the intensity of expressions from onset to peak and to offset state) of video sequences pose additional challenges, making it challenging to recognize the expressions in correlated frame sequences compared to static image analysis. Many recent works, including Renda et al. (2019), have attested that FER on static images is still an active research area. This work focuses on FER based on static images rather than video sequences.

The contributions of the chapter are :

- Different from previous approaches, the top five best weights and their respective intermediate features are fed to the proposed ensemble models. Thus, the frequency-based and stacking classifier approach showed enhanced performance than other existing machine and deep learning techniques.
- Individual machine learning classifiers are assessed using different parameters. A fusion of identical and a diverse set of machine learning classifiers with a frequency-based approach and stacking classifier maintains good efficiency, thus achieving state-of-the-art on posed datasets.
- The proposed model is evaluated on both single and multi-pose datasets with

fine-tuned parameters, making the model achieve better performance against pose variations.

- The proposed model tries to reduce the errors by analyzing the pattern in the base classifier before making the final predictions.

The chapter is organized as follows: Section 5.1 is for introduction, Section 5.2 provides the overview of EfficientNet and Stacking classifier techniques. Section 5.3 describes the proposed model for classification of posed expressions. Section 5.4 presents the results and its detailed analysis using various evaluation parameters, and the chapter is concluded in section 5.5.

## 5.2 PRELIMINARIES

### 5.2.1 EfficientNet

Earlier deep learning models have reached a hardware memory limit issue; hence, an efficient model is required to improve the accuracy. Furthermore, the CNN's are computationally intensive compared to machine learning models, as Neural Networks heavily depend on the data, the problem considered, and the complex network required to solve it. But, the computational difficulty of these networks was solved using Graphical Processing Unit (GPU). Finally, the Google Research Brain team's latest model, the EfficientNet (Tan and Le 2019) (a variant of the CNN), achieved state-of-the-art accuracy, faster computation power, compactness, and overcame all previous deep learning models.

The ConvNets are scaled up to obtain better accuracy and efficiency. Hence, it is scaled up by depth, width, and resolution. Single dimension scaling models tend to achieve higher accuracy with larger depth, width, and resolution, but it has a limitation, the accuracy gain drops and saturates after reaching 80%. The EfficientNet model overcomes the drawback by compound scaling (Tan and Le 2019), i.e., by scaling three dimensions like width, depth, and resolution with a fixed ratio. This model starts from high quality and with a compact baseline model and scales up each of its dimensions uniformly with a fixed set of escalate coefficients. If the image's resolution is bigger in the compound scaling method, the network needs a more



receptive field and more channels to capture other fine-grained patterns. In the proposed work, EfficientNet B0, a baseline model is utilized, and architecture details are given in Table 5.1.

Mobile Inverted Bottleneck Conv (MBConv) (Howard et al. 2019, 2017; Luz et al. 2021; Sandler et al. 2018), an inverted bottleneck Conv, is the main building block or main component of EfficientNet. It is also an inverted residual structure with an injection of Squeeze and Excite (SE) block, which has skip connections between thin bottleneck layers. The inverted residual blocks are efficient compared to classical residual networks, as propagating the gradient across multiplier layers is improved.

Table 5.1: Architecture of EfficientNet B0 (Tan and Le 2019)

Stage $i$	Operator $F_i$	Resolution $H_i \times W_i$	# Channels $C_i$	# Layers $L_i$
1	Conv3×3	224 × 224	32	1
2	MBConv1, k3×3	112 × 112	16	1
3	MBConv6, k3×3	112 × 112	24	2
4	MBConv6, k5×5	56 × 56	40	2
5	MBConv6, k3×3	28 × 28	80	3
6	MBConv6, k5×5	14 × 14	112	3
7	MBConv6, k5×5	14 × 14	192	4
8	MBConv6, k3×3	7 × 7	320	1
9	Conv1×1 & Pooling & FC	7 × 7	1280	1

### 5.2.2 Stacking Classifier

Stacking is a process of constructing classifier ensembles (Aggarwal 2015). It is an ensemble learning technique that combines multiple classification models (machine learning classifiers) via meta classifier (Malmasi and Dras 2018; Sakkis et al. 2001). It is an approach where several individual classifiers' outputs (decisions) are combined to classify new instances. The stacking process combines multiple classifiers (Li and Zou 2017; Mihalcea 2002) to create high-level classifiers and produce improved performance. In the first level, the features are fed into the various base classifiers

which, outputs a new decision. Later in a second level, a meta classifier decides the final prediction by considering the base classifiers' opinions and their prediction (output pattern) value (Álvarez et al. 2016). Suppose if the base classifiers make some classification errors. In that case, the meta classifier can successfully learn the pattern and decide which prediction value to be considered for the final prediction. By doing so, the overall performance of the recognition system is improved. The bias and variance can be reduced using the stacking approach (Aggarwal 2015), as the combination of different ensemble components tries to learn from its errors. The stacking approach is flexible and powerful as compared to that of other ensemble methods.

### 5.3 PROPOSED MODEL FOR CLASSIFICATION OF POSED EXPRESSIONS

The proposed model consists of three stages: Pre-processing, Feature Extraction, and Classification. Finally, the facial images are mapped into respective expression classes using an ensemble approach as shown in Figures 5.1 and 5.2. The proposed model's performance is evaluated on Oulu-CASIA and RaFD (multi-pose and only frontal pose images) datasets.

#### 1. Pre-processing and Data Augmentation

The static images are used to carry out the experiments. Every image present in the dataset is pre-processed. For the face detection, Paul Viola and Michael Jones (Viola and Jones 2001, 2004) Adaboost learning algorithm is used. This technique uses Haar-Like features and AdaBoost to train cascaded classifiers and detect the faces with a frontal view in lesser time (Zia et al. 2018). Facial image contains a lot of unnecessary background information which is not useful for the classification of the expressions (Lopes et al. 2017), hence this irrelevant information is cropped, and expression specific information is retained. This approach was successful on Oulu-CASIA and RaFD datasets with only a frontal pose. The Viola-Jones algorithm failed to detect the faces with pose variations (Zhang et al. 2016). Hence, a MTCNN is used to detect faces with multi-pose variations. MTCNN (Zhang et al. 2016) is a deep cascaded architecture which

### 5.3. Proposed model for classification of Posed Expressions

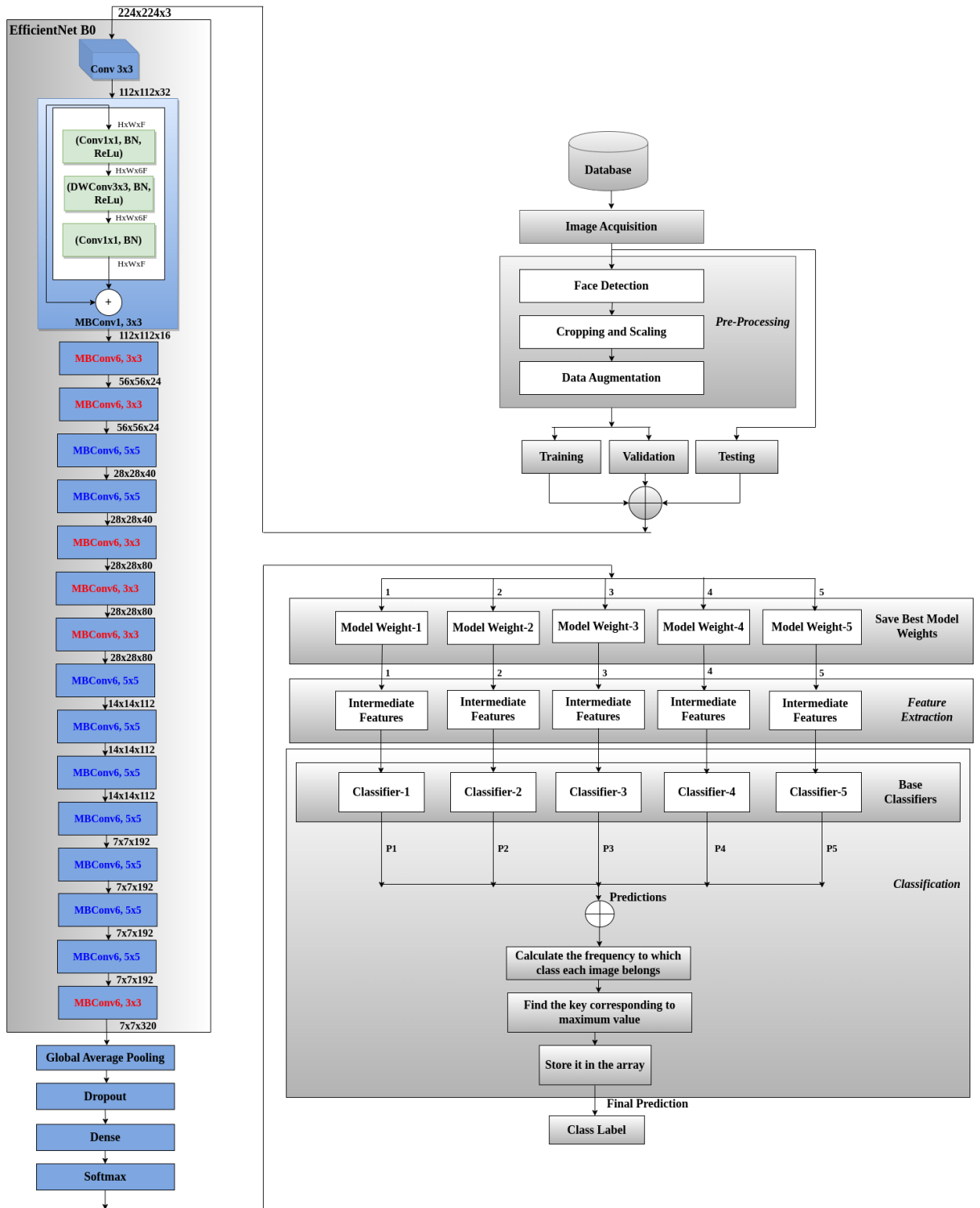


Figure 5.1: Ensemble model architecture based on the frequency of votes (FV-EffNet).

## 5. Classification of Posed Expressions

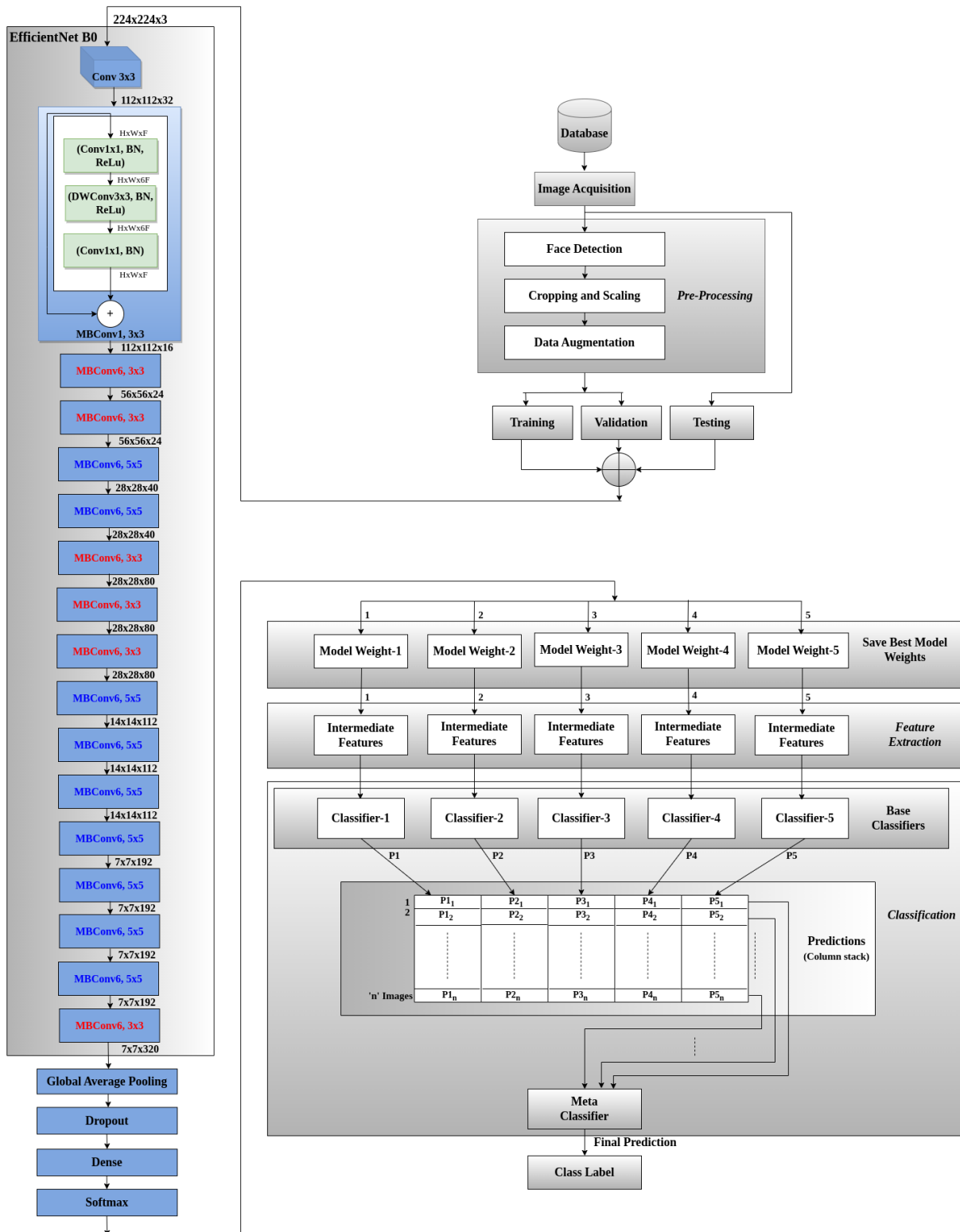


Figure 5.2: Stacking classifier architecture (SC-EffNet)

exploits the innate correlation between the detection and alignment. This framework consists of three-stage multitask deep convolutional networks like Proposal Network (P-Net), Refine Network (R-Net), and Output Network (O-Net), designed to predict facial and landmark location in a coarse-to-fine manner. The MTCNN face detection technique was successful in detecting faces with a multi-pose variation on the RaFD dataset. All the images present in the dataset were resized into 224x224 pixel resolution and fed to the network for recognition.

The offline data augmentation is performed to improve the training samples. Data augmentation is a technique to virtually create extra training data by applying transformations to the input image, given the training data. In this work, horizontal and vertical flipping is applied on Oulu-CASIA and RaFD (frontal data) datasets. Data Augmentation has proven to be efficient in improving the generalization ability of deep learning models in various applications like image classification, speech recognition, and other areas. The huge complex designed network tend to over-fit on training data. Hence, to avoid this, it requires to feed a massive amount of data (Kuang et al. 2016).

## 2. Feature Extraction using deep learning Model

Feature extraction is a vital step in FER, and it is an aid to derive effective facial representations from the original facial image. There are two ways of extracting facial features: handcrafted features and the other using CNN architectures to extract auto-extracted features (Tang et al. 2018). The extracted features play an essential role in minimizing the distance of intra-class variations and maximizing the distance between inter-class. The best classifier will also fail to achieve good performance if the extracted features are inadequate. The handcrafted feature extraction techniques like LBPs, LGBPs, HOG, and SIFTs have achieved great success in various fields with a small amount of training data. These low-level features are difficult to extract; tuning the features according to incoming face images and gathering discriminative information from these data is also tricky. These disadvantages present significant challenges

in accurately recognizing expressions in real-life applications, as these data impose large inter-personal differences in appearance and capturing conditions. Hence, deep learning approaches cope up with these challenges and automatically discover multiple data representations and extract abstract concepts from a higher representation level. Thus, this was a reason for the breakthrough in recognition tasks.

The EfficientNet model has achieved a state-of-the-art in image classification (Tan and Le 2019) and achieved high performance, and low computational cost (Luz et al. 2021). In the proposed work, the EfficientNet B0 architecture has been used in the basic feature extraction process. Initially, the EfficientNet model is executed up to certain epochs, and all the weights are saved into a folder. Then, the model's five best weights are chosen based on the performance of the validation accuracy. Later, the best weights are loaded, and their respective intermediate features are extracted and fed into an ensemble of machine learning classifiers to improve the FER's efficiency.

### 3. Classification

Combining the EfficientNet model's features and various machine learning classifiers proved to be advantageous (Malmasi and Dras 2018). This work presented two novel ensemble models, an EfficientNet model with machine learning classifiers using a frequency-based voting strategy (FV-EffNet) and an EfficientNet model with the stacking classifier (SC-EffNet) for classification. Various machine learning classifiers are evaluated empirically. The classifiers that generated the best results were chosen for further evaluation. The proposed approaches takes machine learning classifiers like Extra Trees classifier (Yaddaden et al. 2018), RF (Aggarwal 2015), DTree (Safavian and Landgrebe 1991), KNN (Dino and Abdulrazzaq 2019), MLP (Rashid 2016), and SVM (Michel and El Kaliouby 2003) as base classifiers.

Every model's intermediate features are loaded individually and fed into each base classifier separately and evaluated to check with which intermediate features and base classifier (varying their parameters) the efficient result is

obtained. This step is necessary as the features fed to the base classifier play a vital role in recognizing the input pattern and predicting the outcomes.

(a) Classification using a frequency-based voting approach: In FV-EffNet for each image, the predictions from five separate base classifiers are analyzed row by row (predicting the class to which each image belongs). The frequency (vote) is calculated using all five predictions for each image. Finally, the maximum vote from all five predictions is used to generate a key. The final key value is the final prediction (emotion class to which each image belongs) obtained from the combination of base classifiers, and it is stored in the array. The strategy is depicted in the Figure [5.1](#).

- Base classifiers: The EfficientNet B0 intermediate features are fed to various machine learning classifiers. The base classifiers like KNN, MLP, RF, Extra Trees, SVM on Oulu-CASIA, and Extra Trees, RF, DTree, MLP, KNN on RaFD datasets are chosen to predict various expression classes efficiently.
- Ensemble classifier: The predictions from the aggregation of base classifiers would outperform compared to predictions from a single model ([Malmasi and Dras 2018](#)). Hence, this is a reason behind the choice of an ensemble model to predict expression class. A frequency-based voting strategy combines the predictions from various machine learning base classifiers and makes a final decision. Hence, the ensemble model's output using a frequency-based algorithm would be a final class label predicted by most classifiers ([Rao et al. 2019](#)).

(b) Classification using stacking classifier: It is an ensemble learning technique that combines multiple classification models through meta-classifier ([Malmasi and Dras 2018](#)). Instead of bagging and boosting approach, stacking tries to learn how to combine the base classifiers (first-level classifiers) rather than taking votes. A novel approach, where a combination of deep learning model and stacking classifier (SC-EffNet) is proposed and depicted in Figure [5.2](#). The best features from the

EfficientNet B0 model are fed to a stacking classifier (Aggarwal 2015; Tang et al. 2015) for classification.

The intermediate features are first fed to base classifiers (level 0 classifiers), containing diverse machine learning classifiers. Each base classifier is trained using training data. The predictions (output) from each base classifier are appended and stacked as a vector. These predictions are considered as a new dataset fed as input to the meta classifier (level 1 classifier) (Aggarwal 2015). Later, the meta classifier is trained with this new dataset, and evaluation is done by performing cross-validation on test data. The meta classifier helps analyze the data pattern in a better way and helps to get accurate predictions. Finally, this classifier outputs the final prediction. One of the advantages of using a stacking classifier is that it decreases the risk of getting varied outputs from different machine learning classifiers. It clubs the results of all individual machine learning classifiers, analyzes the pattern, performs accurate predictions, and achieves good performance.

- Base Classifiers: The machine learning classifiers like KNN, MLP, RF, Extra Trees, and SVM are used as a base classifier on the Oulu-CASIA dataset and Extra Trees, RF, DTree, MLP, KNN on the RaFD dataset.
- Meta Classifier: The Extra Trees classifier outperformed other machine learning classifiers and was chosen as a meta classifier for evaluation on the Oulu-CASIA dataset. During the RaFD dataset evaluation, the DTree classifier proved to be efficient compared to other machine learning classifiers for the final prediction.

## 5.4 RESULTS AND DISCUSSIONS

### 5.4.1 Dataset Description

This section presents different datasets used for the evaluation of the proposed technique.



1. Oulu-CASIA: The images captured from visual cameras are used for the evaluation of the proposed methodology. In this experiment, peak expressions from the last three frames are chosen as training, testing, and validation data. A total of 236 images have been chosen for our experiments from 240 image segments after applying the viola jones face detection algorithm from all expression classes. Images are resized into 224\*224 pixel resolution. Later, horizontal and vertical scaling augmentations are applied to increase the number of images. This dataset includes basic emotions like anger, disgust, fear, happiness, sadness, and surprise (Zhao et al. 2011).
2. Radboud Faces Database (RaFD): Two types of experimentation are carried out in this work using the RaFD database. One experiment is carried out with images with only a frontal pose, where the Viola-Jones algorithm is applied for face detection. The augmentation like horizontal and vertical scaling is used to increase the images with a frontal pose. The other experiment is carried out with the entire RaFD dataset (Langner et al. 2010; Shokrani et al. 2014), which includes all five pose angles (Pose degree: 0, 45, 90, 135, 180 with three gaze directions (frontal, left and right views)) as depicted in the Figure 5.3. The dataset includes expressions like anger, disgust, fear, happiness, sadness, surprise, contempt, and neutral. A total of 7974 facial images are detected using the MTCNN face detection approach. Data augmentation is not applied to this data with multi-pose angle, and all images are resized into 224\*224 pixel resolution.



Figure 5.3: The five pose angles with cameras in the order 180°, 135°, 90°, 45° and 0°

### 5.4.2 Implementation Details

For training the network, this work has used a pre-trained network with pre-trained weights instead of training from scratch, and this is known as transfer learning. Transfer learning has proved to be effective in various computer vision applications (Luz et al. 2021). This approach is applied to EfficientNet B0 and pre-trained on the ImageNet dataset, much broader than the facial images presented to the proposed models. The network weights are fine-tuned by the optimizer in the new training phase, allowing the model to adapt to our problem. The imported models have a lot of knowledge about the objects.

In the training phase, Adam optimizer (Zhang 2018) is used to update the weights and reduce the learning rate by a factor of 10 in the event of stagnation ('patience=7'). The learning rate started at  $1e^{-4}$ , and the batch size is 10, and the number of epochs is fixed at 50. During training, an early stopping callback is used to control the overlearning of EfficientNet architecture. The model weights are saved using the model checkpoint. ReLU (Verma et al. 2019) activation function is used which transforms the linear input into non-linear data. ReLU is computed using formula  $f(\varphi) = \max(0, \varphi)$ . With ReLU, the network becomes more efficient due to its sparse feature representation; it also helps in faster training, reduces computational complexity, and overcomes vanishing gradient problem. The softmax layer is used in the output layer of the EfficientNet model. It is used in multi-class classification problems (Renda et al. 2019) to estimate the testing sample's probabilities belonging to each class.

### 5.4.3 Experiment 1: Evaluation of EfficientNet B0 model

In this experiment, the EfficientNet B0 model is used as a classifier for evaluating posed datasets, and the results are presented in Table 5.2. The model showed an accuracy of 97.28% and 98.53% with augmented data on Oulu-CASIA and RaFD (only 90 deg frontal pose) datasets. Without data augmentation, the EfficientNet B0 model as a classifier achieved a performance of 93.72%, 95.10%, and 97.06% on Oulu-CASIA, RaFD datasets with frontal pose and Multi-pose variations, respectively.

Table 5.2: EfficientNet B0 Architecture results when utilized as a Classifier

Model	Optimizer	Dataset	Test Accuracy
EfficientNet B0	Adam	Oulu-CASIA (without augmentation)	93.72%
		Oulu-CASIA (with augmentation)	97.28%
		RaFD (90 deg) (without augmentation)	95.10%
		RaFD (90 deg) (with augmentation)	98.53%
		RaFD (Multi-Pose) (without augmentation)	97.06%

#### 5.4.4 Experiment 2: Proposed Methodology

The entire dataset is split into training, testing, and validation set. Every epoch's weights are saved while monitoring the parameter `val_acc` using the model checkpoint. An early stopping callback is used to stop the EfficientNet model's training if there is no increase in the value of `val_acc` until `patience=10` (10 iterations). Among all the saved weights, 'n' best weights are loaded where `n=1` to 5. Their respective intermediate features are fed to the combination of machine learning base classifiers for classification as shown in Figures 5.1 and 5.2.

The detailed experiment procedure outlining the entire flow of the proposed methodology is depicted in Figure 5.4. First, according to the process, the model weights that achieved the best results are saved, and their respective intermediate features are loaded. Next, every machine learning classifier is adapted, and these classifiers' performance is verified on the saved 'n' models intermediate features. Finally, the classifier which gave the best results is chosen as a particular base classifier for that specific model. Similarly, the same procedure is followed for the rest of the 'n-1' models. The parameters of the machine learning classifiers are fine-tuned by varying the number of estimators, maximum depth of the tree, minimum samples of a leaf node, minimum sample split, maximum iterations of RF, DT, Extra Trees, SVM, MLP classifiers, and the nearest neighbor value of the KNN classifier. The parameters that showed the best performance results were eventually considered for individual classifiers. The Figures 5.5 and 5.6 presents results obtained from the base classifiers

## 5. Classification of Posed Expressions

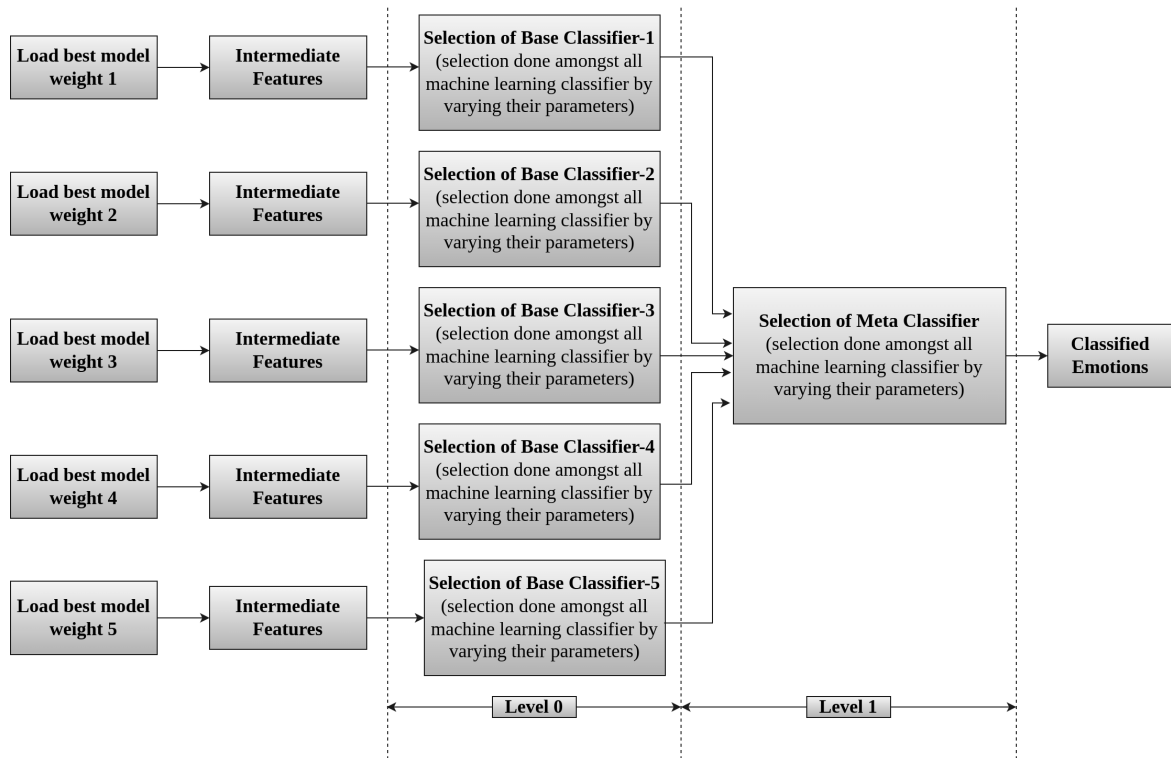


Figure 5.4: Detailed procedure outlining the flow of proposed methodology

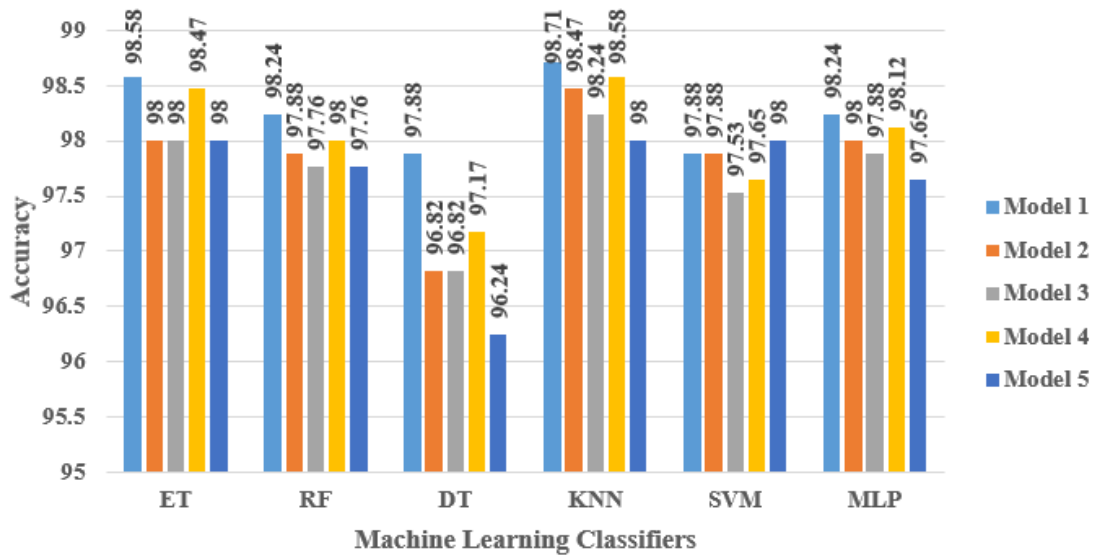


Figure 5.5: Results obtained from base classifiers (level 0) on Oulu-CASIA dataset

when evaluated using identical set of machine learning classifiers.

### a. Frequency-based Voting Strategy:

Five best epochs intermediate features are fed to ‘m’ machine learning base classifiers where  $m=1$  to 5. The predictions from five individual machine

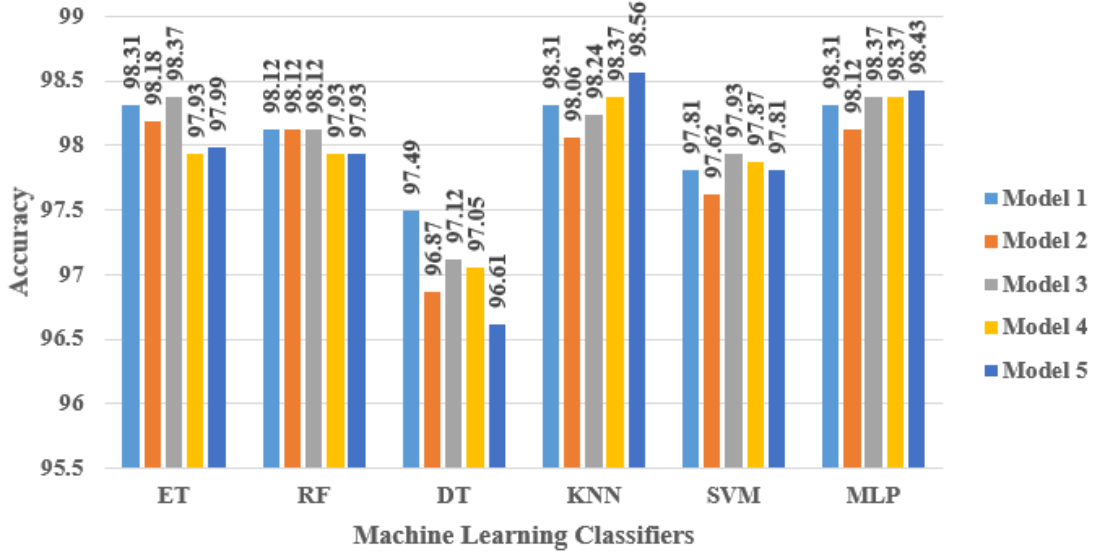


Figure 5.6: Results obtained from base classifiers (level 0) on RaFD dataset (Multi-Pose)

learning classifiers (base classifiers) are appended and considered for further evaluation. The final class label is predicted by taking votes from most classifiers in the ensemble model. The experiment results obtained from the proposed approach are presented in Tables 5.3 and 5.4. The deep learning model (EfficientNet B0), combined with distinct classifiers using a frequency-based voting strategy, gave the best results of 98.71% and 98.56% on Oulu-CASIA and RaFD multi-pose datasets, respectively and also depicted in Figures 5.8 and 5.9.

#### b. Stacking Classifier:

Initially, five best epochs intermediate features are fed to ‘m’ machine learning classifiers (base classifiers) where  $m=1$  to 5. After empirically testing each machine learning classifier, the classifiers that performed best are chosen as the base classifier. All the base classifiers are trained using training data, and their predictions are horizontally stacked and converted into vectors and fed to meta classifier to get the final prediction. Due to a lack of test data, the test set is subjected to Cross-Validation ( $CV=5$ ). As a result, it establishes the robustness of the stacking strategy and the model’s generalizability. Each machine learning classifier (Extra Trees, KNN, RF, DTree, MLP, and SVM) is individually chosen

## 5. Classification of Posed Expressions

Table 5.3: Results of fine-tuning individual classifiers on Oulu-CASIA and RaFD datasets.

Model	Base Classifiers	Meta Classifier	Oulu-CASIA			RaFD (Multi-Pose)		
			Output	Voting	Stacking Classifier	Output	Voting	Stacking Classifier
EfficientNet B0	Extra Trees 1	Extra Trees	98.58%	98.35%	98.47%	98.31%	98.24%	98.18%
	Extra Trees 2		98%			98.18%		
	Extra Trees 3		98%			98.37%		
	Extra Trees 4		98.47%			97.93%		
	Extra Trees 5		98%			97.99%		
	RF 1	RF	98.24%	98.12%	98%	98.12%	98.24%	98.18%
	RF 2		97.88%			98.12%		
	RF 3		97.76%			98.12%		
	RF 4		98%			97.93%		
	RF 5		97.76%			97.93%		
	DTree 1	DTree	97.88%	98.24%	97.65%	97.49%	98.31%	96.74%
	DTree 2		96.82%			96.87%		
	DTree 3		96.82%			97.12%		
	DTree 4		97.17%			97.05%		
	DTree 5		96.24%			96.61%		
	KNN 1	KNN	98.71%	98.47%	98.24%	98.31%	98.18%	97.93%
	KNN 2		98.47%			98.06%		
	KNN 3		98.24%			98.24%		
	KNN 4		98.58%			98.37%		
	KNN 5		98%			98.56%		
	SVM 1	SVM	97.88%	98.35%	97.83%	97.81%	98.24%	96.31%
	SVM 2		97.88%			97.62%		
	SVM 3		97.53%			97.93%		
	SVM 4		97.65%			97.87%		
	SVM 5		98%			97.81%		
MLP 1	MLP	98.24%	98.47%	98.24%	98.31%	98.37%	97.93%	
MLP 2		98%			98.12%			
MLP 3		97.88%			98.37%			
MLP 4		98.12%			98.37%			
MLP 5		97.65%			98.43%			

\* Meta classifier is considered only for the evaluation of stacking classifier approach.

for evaluation as a meta classifier. Further, as shown in Figure 5.7, based on various meta classifiers' performances, Extra Trees and DTree classifiers proved to be efficient on Oulu-CASIA and RaFD datasets.

The results obtained when evaluating the identical base classifier and the same

Table 5.4: The output of distinct machine learning algorithms after fine-tuning on Oulu-CASIA and RaFD datasets.

Datasets	Base Classifiers	Meta Classifier	Output	Voting	Stacking Classifier
Oulu-CASIA	KNN	Extra Trees	98.71%	<b>98.71%</b>	<b>98.35%</b>
	MLP		98%		
	RF		97.76%		
	Extra Trees		98.47%		
	SVM		97.88%		
RaFD (Multi Pose)	Extra Trees	DTree	98.31%	<b>98.56%</b>	<b>98.06%</b>
	RF		98.12%		
	DTree		97.12%		
	MLP		98.37%		
	KNN		98.56%		

\* Meta classifier is considered only for the evaluation of stacking classifier approach.

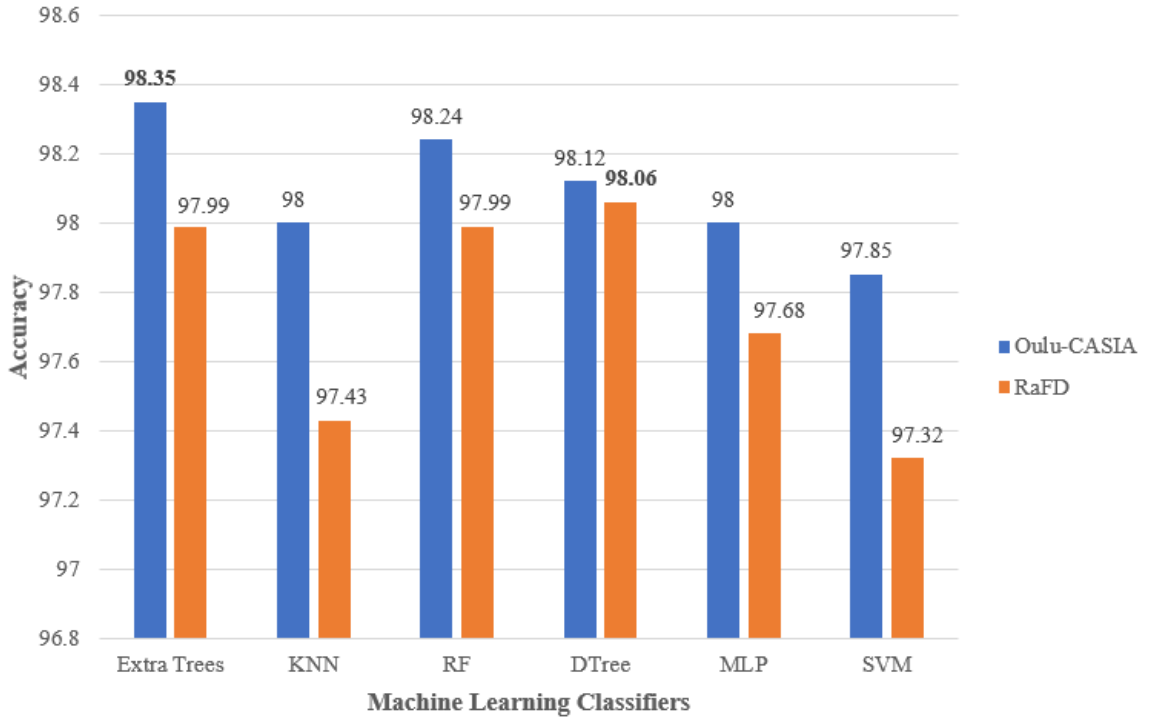


Figure 5.7: Selection of meta classifier (level 1) for evaluation of distinct set of base classifiers in stacking classifier approach.

machine learning classifier chosen as meta-classifiers are depicted in Figures 5.8 and 5.9 and also in Table 5.3. Similarly, the EfficientNet model with a distinct

## 5. Classification of Posed Expressions

combination of machine learning classifiers and stacking classifiers are presented in Figures 5.8 and 5.9 and also in Table 5.4. The accuracy of 98.35% and 98.06% is obtained with a stacking classifier approach on Oulu-CASIA and RaFD (multi-pose) datasets.

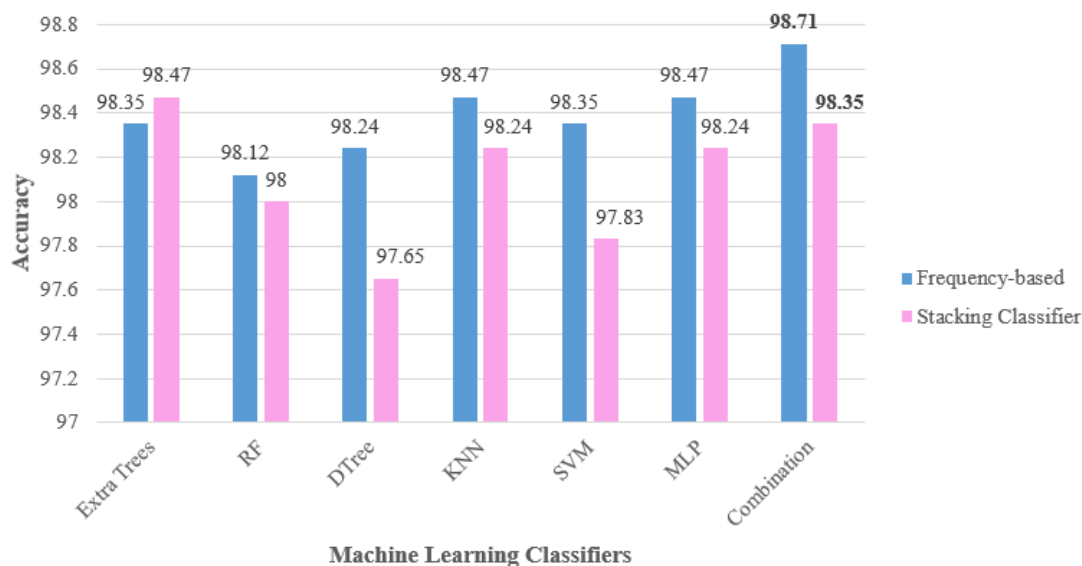


Figure 5.8: Results obtained when evaluating the frequency-based and stacking classifier approaches on Oulu-CASIA dataset.

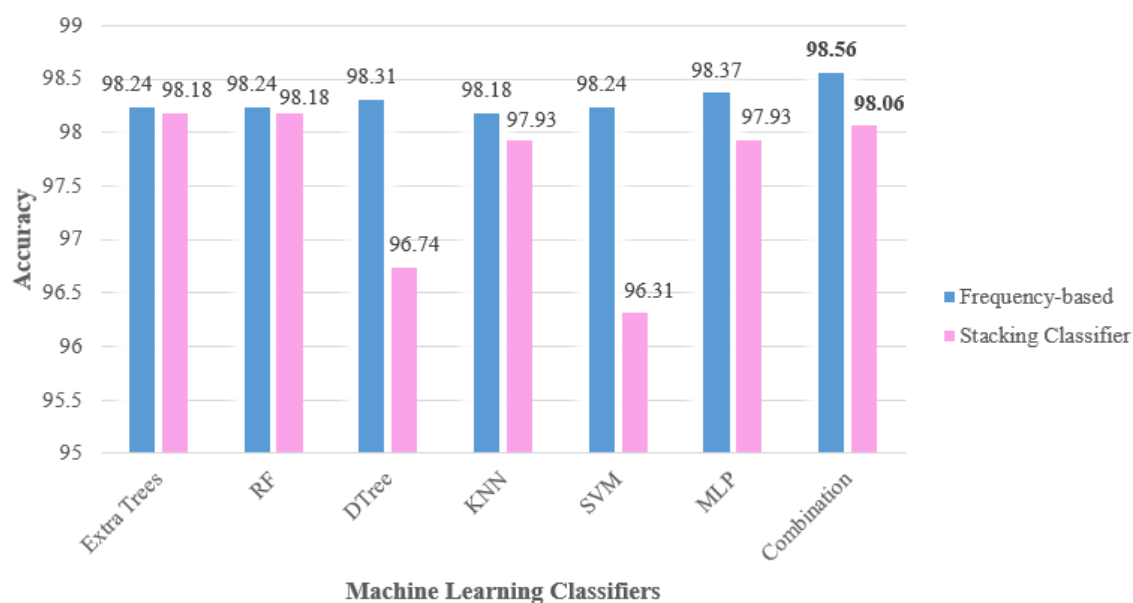


Figure 5.9: Results obtained when evaluating the frequency-based and stacking classifier approaches on RaFD dataset (Multi-Pose).



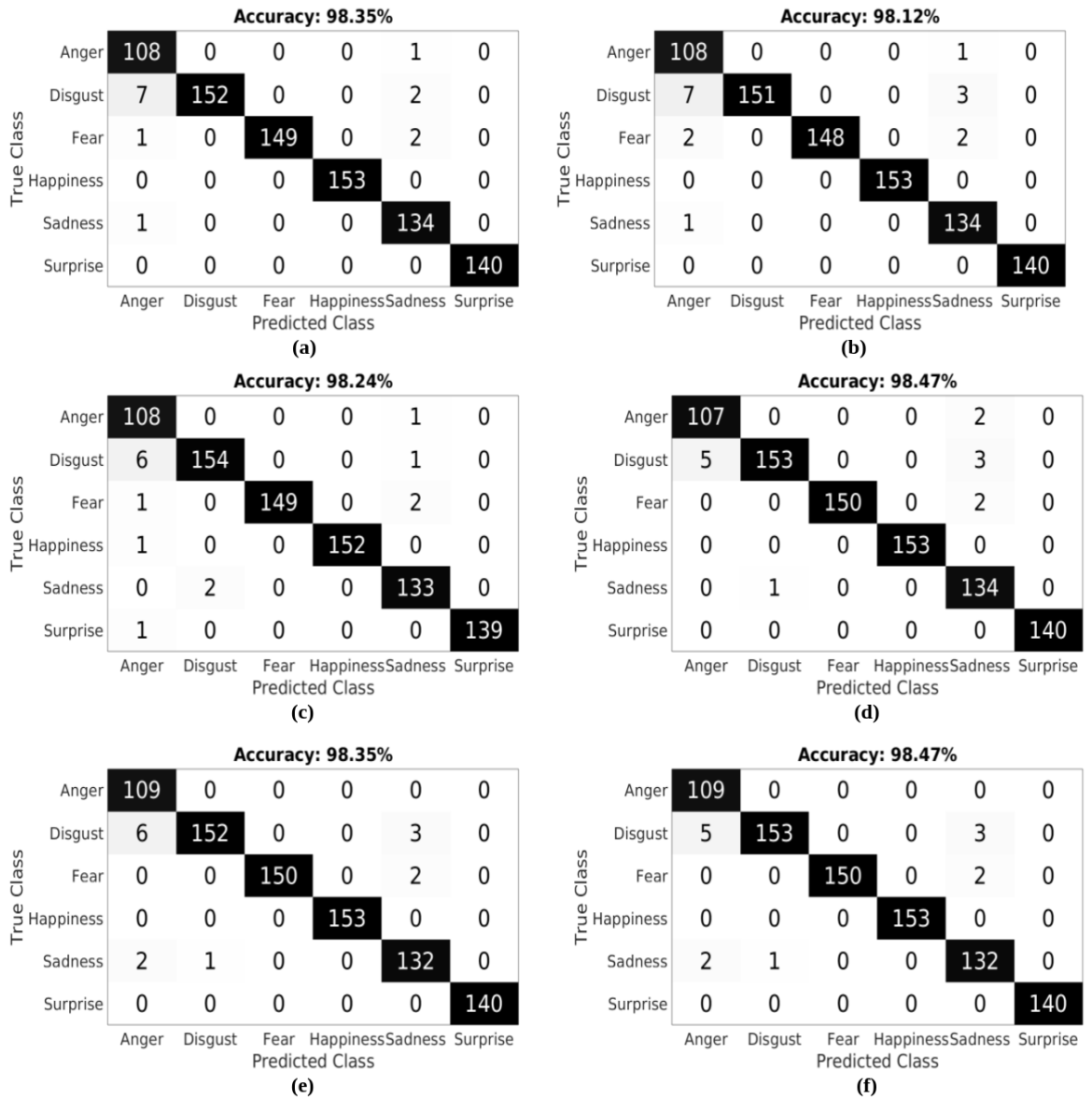


Figure 5.10: Confusion Matrices obtained from machine learning classifiers on Oulu-CASIA dataset ((a) Extra Trees Classifier (b) Random Forest (RF) (c) Decision Trees (DTree) (d) K-Nearest Neighbors (KNN) (e) Support Vector Machine (SVM) (f) Multi-Layer Perceptron (MLP))

### 5.4.5 Observations

The confusion matrices obtained when evaluating a combination of individual machine learning classifiers are presented in Figures 5.10 and 5.11. Also, the Figures 5.12 and 5.13 depicts the confusion matrices obtained when evaluating a different combination of machine learning classifiers on Oulu-CASIA and RaFD (multi-pose) datasets, respectively. Thus, the observation is that every individual classifier contributes

## 5. Classification of Posed Expressions

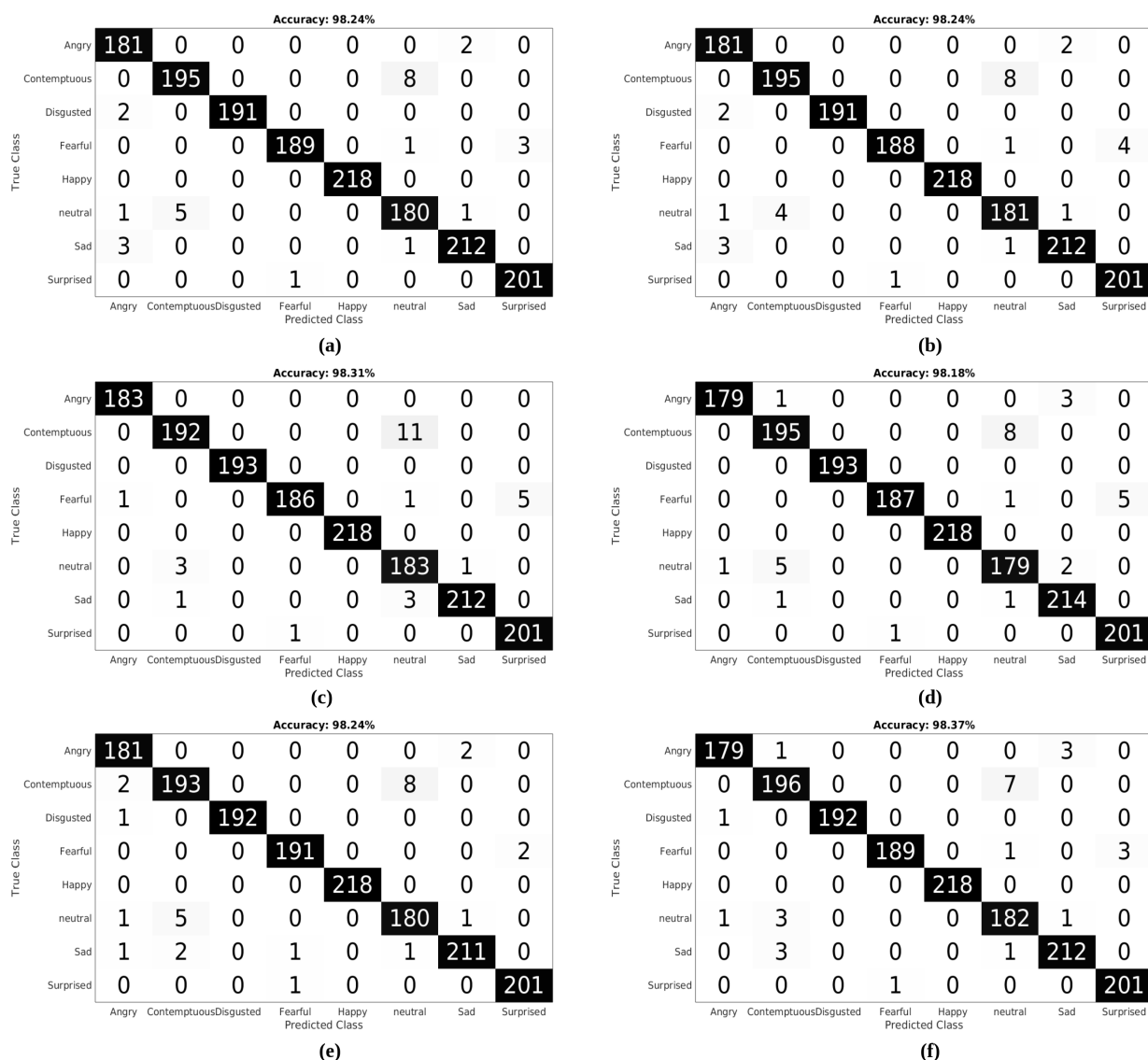


Figure 5.11: Confusion Matrices obtained from machine learning classifiers on RaFD (Multi-Pose) dataset ((a) Extra Trees Classifier (b) Random Forest (RF) (c) Decision Trees (DTree) (d) K-Nearest Neighbors (KNN) (e) Support Vector Machine (SVM) (f) Multi-Layer Perceptron (MLP))

equivalently to classify the expressions into respective classes. For example, while observing the confusion matrix given in Figure 5.10, the Extra Trees, RF, and DTree classifiers predict 108 images correctly into anger expression classes and does one misclassification into a sadness expression class. Similarly, the KNN classifier predicts 107 images correctly into anger expression class and does two misclassifications. Whereas SVM and MLP classifier precisely classifies all 109 images into proper expression classes. Thus, the conceptual lesson to be learned from this work is that

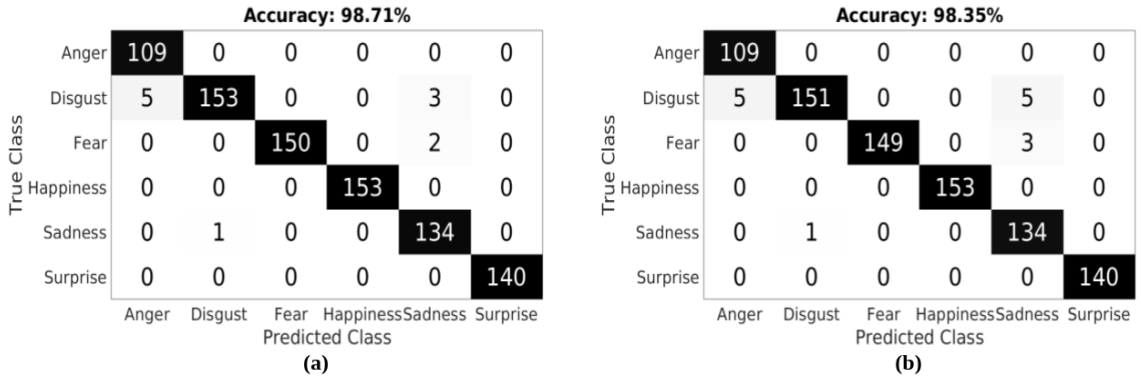


Figure 5.12: Confusion Matrix obtained from the combination of various machine learning classifiers on Oulu-CASIA dataset (a) FV-EffNet (b) SC-EffNet

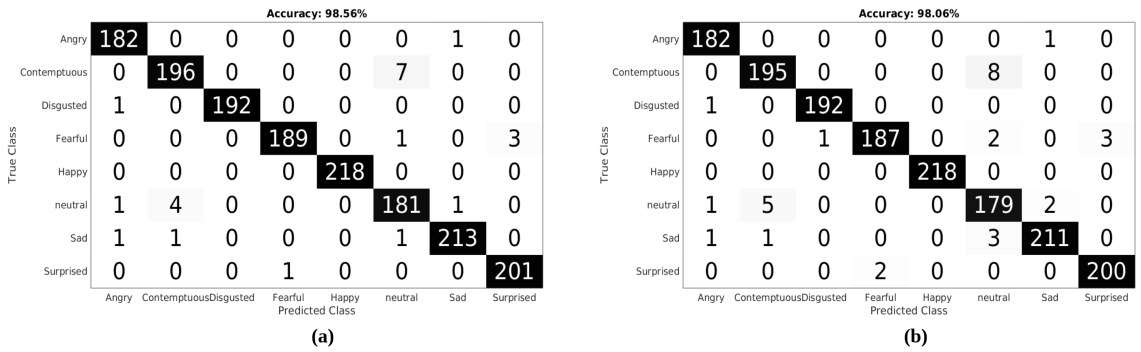


Figure 5.13: Confusion Matrix obtained from the combination of various machine learning classifiers on RaFD (Multi-Pose) dataset

every individual classifier is responsible for categorizing expressions into an appropriate emotion class.

The best features fed into an ensemble of machine learning classifiers showed enhanced performance on the frontal pose and multi-pose datasets. By fusing the outputs of base classifiers and providing them to the higher-level classifier, we try to reduce the errors by analyzing the pattern before making the final predictions. Thus, it suggests that a combination of classifiers using the stacking approach is better than selecting the best single classifier for classification. It will help improve the system’s efficiency and overcome the mistakes made in the previous classification level. The ensemble of deep learning and machine learning techniques performs better than the earlier methods, thus showing the state-of-the-art on Oulu-CASIA and RaFD datasets. The proposed approach is robust against pose variations and involves multiple processing stages. However, majority voting predominately aids in enhancing the

effectiveness of the system.

### 5.4.6 Experiment Analysis and Comparisons

The experiment results obtained with previous FER studies are presented in Table 5.5 and compared with the proposed approach. With the proposed methodology, the accuracies of 98.71% and 98.56% using frequency-based strategy and 98.35% and 98.06% using a stacking classifier approach are obtained on Oulu-CASIA and RaFD multi-pose datasets, respectively. The performance of the proposed approach is compared with other machine learning methods, CNN-based methods, and state-of-the-art results on the Oulu-CASIA and RaFD datasets. As observed in Table 5.5, the proposed model achieves better results than other methods on these benchmark facial expression databases.

Using RaFD datasets with a frontal pose, the proposed model outperforms Sun et al. (2019) and Happy et al. (2019) by 0.83% and 1.42%, respectively. The authors used a pyramid histogram of an oriented gradient for feature extraction with KNN classification in Shokrani et al. (2014) and attained 100% accuracy. With an ensemble model, the proposed method for the RaFD dataset with the frontal pose obtained 100% accuracy. The intermediate features extracted from the best weights of the EfficientNet model tried to analyze the pattern and perform precise predictions for the RaFD dataset with 90-degree pose variation. Also, the facial images were clearly visible. Hence, the proposed model extracted the efficient features and reduced the risk of getting varying output before proceeding to the meta-classifier for further classification. Additionally, eight expression classes were considered in the proposed work, which improved the accuracy rate to 2.29%, outperforming Wu and Lin (2018). The authors in Wu and Lin (2018) avoided the contemptuous class and employed seven expression classes. On the Oulu-CASIA dataset, the proposed technique had an enhanced accuracy of 10.71% compared to Yang et al. (2018), which extracted the expressive component through a deexpression mechanism.

Table 5.6 presents the experiment results evaluated using various other performance metrics. When using the RaFD dataset with a frontal pose, the proposed

Table 5.5: Comparison with previous approaches on Oulu-CASIA and RaFD datasets

Dataset	Method	Experimental Settings	Accuracy
Oulu CASIA	DTAGN (Joint) (Jung et al. 2015)	Sequence-based	81.46%
	STM-ExpLet (Liu et al. 2016)	Sequence-based	74.59%
	DFSN-I (Tang et al. 2018)	Sequence-based	87.50%
	PHRNN-MSCNN (Zhang et al. 2017)	Sequence-based	86.25%
	Microexpnet (Cugu et al. 2019)	Sequence-based	95.02%
	PPDN (Zhao et al. 2016)	Image-based	84.59%
	DeRL (Yang et al. 2018)	Image-based	88%
	FN2EN (Ding et al. 2017)	Image-based	87.71%
	EfficientNet B0 (Baseline)	Image-based	97.28%
	Proposed methodology using majority voting	Image-based	<b>98.71%</b>
	Proposed methodology using stacking classifier	Image-based	<b>98.35%</b>
RaFD (Frontal pose)	18-layered Conv-Deconv (Wenyun et al. 2018)	Image-based	93.41%
	CNN (Fathallah et al. 2017)	Image-based	93.33%
	MDSTFN (Sun et al. 2019)	Image-based	99.17%
	CNN (González-Hernández et al. 2018)	Image-based	95%
	Weakly supervised learning (Happy et al. 2019)	Image-based	98.58%
	Hybrid-based AFER (Yaddaden et al. 2018)	Image-based	96.16%
	Metric Learning (Jiang and Jia 2016)	Image-based	95.95%
	BAE-BNN-3 (Sun et al. 2017)	Image-based	96.93%
	HOG-SVM (Carcagni et al. 2015)	Image-based	98.2%
	EfficientNet B0 (Baseline)	Image-based	98.53%
	Proposed methodology using majority voting	Image-based	<b>100%</b>
Proposed methodology using stacking classifier	Image-based	<b>100%</b>	
RaFD (Multi Pose)	MP-AdaBoost (Jiang and Jia 2013)	Image-based	82.68%
	SURF boosting (Rao et al. 2015)	Image-based	90.64%
	W-CR-AFM (Wu and Lin 2018)	Image-based	96.27%
	PHOG-KNN (Shokrani et al. 2014)	Image-based	100% (90deg), 96.7% (45deg to the right) and 98.1% (45deg to the left)
	Semi-supervised DBN (Kurup et al. 2019)	Image-based	91.95% (135deg), 94.50% (90deg) and 92.75% (45deg)
	EfficientNet B0 (Baseline)	Image-based	97.06%
	Proposed methodology using majority voting	Image-based	<b>98.56%</b>
Proposed methodology using stacking classifier	Image-based	<b>98.06%</b>	

model performs better in terms of precision, recall, and F1-score than Carcagni et al. (2015). Before making the final predictions, the proposed model examines the pattern in the base classifier to minimize the errors. Hence, the results showed better performance compared to previous studies making the proposed system robust against

Table 5.6: Experiment results using other performance metrics

Dataset	Method	Experimental Settings	Precision	Recall	F1-score
RaFD (Frontal pose)	HOG-SVM (Carcagni et al. 2015)	Image-based	93%	92.9%	92.9%
RaFD (Frontal pose)	Proposed methodology	Image-based	100%	100%	100%
RaFD (Multi pose)	Proposed methodology (FV-EffNet)	Image-based	98.53%	98.54%	98.53%
RaFD (Multi pose)	Proposed methodology (SC-EffNet)	Image-based	98.02%	98.04%	98.02%
Oulu-CASIA	Proposed methodology (FV-EffNet)	Image-based	98.56%	98.83%	98.67%
Oulu-CASIA	Proposed methodology (SC-EffNet)	Image-based	98.22%	98.51%	98.33%

pose variations.

## 5.5 SUMMARY

In this chapter, the top best features from EfficientNet B0 models are extracted. The two strategies, such as stacking classifier and majority voting, are proposed to enhance the classification results performance to detect the posed expressions. The proposed frequency-based voting approach (FV-EffNet) and stacking classifier approach (SC-EffNet) deal with profile and frontal pose variations. The combination of multiple base classifiers in the ensemble model induces the higher-level classifier to learn the pattern and make accurate predictions during classification.

In this chapter, a ensemble model with a frequency-based voting approach (FV-EffNet) and a stacking classifier approach (SC-EffNet) is adopted to classify the posed expressions into respective classes. Combining multiple base classifiers induces the

higher level classifiers to learn the pattern and thus help the ensemble model make accurate predictions rather than selecting a single classifier for classification. In the proposed methodology, even though both the ensemble models gave the best results, majority voting predominantly helped to improve the system's performance.

The following conclusions are drawn from the experimental results:

1. The selection of best model weights and features extracted from EfficientNet showed better performance of 98.71% and 98.35% on Oulu-CASIA, and 98.56% and 98.06% on RaFD (multi-pose) datasets using frequency-based and stacking classifier approach compared to a baseline model.
2. An ensemble model with a frequency-based approach showed an improvement of 10.71% on the Oulu-CASIA dataset compared to [Ding et al. \(2017\)](#) and 2.29% on the RaFD multi-pose dataset achieving the best performance than [Wu and Lin \(2018\)](#).
3. The stacking classifier approach showed an improved efficiency by 10.35% and 1.79% on Oulu-CASIA and RaFD datasets, respectively. This method decreases the risk of getting varying results from various machine learning classifiers and reduces bias and variance. The cross-validation performed on the test set proved the model's robustness and generalizability.
4. The proposed method with multi-stage processing showed better results with pose variations.

The future work would be to evaluate the proposed approach on the spontaneous and in-the-wild databases and build a fully automated system that could be feasible for deploying real-world applications. Chapter 6 discusses the implementation details for the recognition of facial engagement levels of students in the MOOC scenarios.





## CHAPTER 6

### FACIAL ENGAGEMENT ANALYSIS

This chapter briefly explains the experiment used to analyze the levels of facial engagement. The facial engagement recognition framework is proposed to assess students' engagement levels in the MOOC scenario.

#### 6.1 INTRODUCTION

Engagement is one of the essential topics in the field of educational psychology. The engagement serves as a link between the student and the learning resource (Shen et al. 2021). It is one of the keys to overcome educational problems like reducing the dropout rates and addressing low achievements of students in academics (Sinatra et al. 2015). Student Engagement (Nkomo et al. 2021) can lead to the development of students thinking skills and aid in retention, thus influence the learning process. Engagement recognition depends on the subject's emotion, movement, and other behavioral features, and it varies with both researcher's theoretical perspectives and the level at which the engagement is conceptualized, observed, and measured. This type of engagement varies from micro to macro level (Sinatra et al. 2015). School engagement can be characterized as a multi-dimensional construct that includes emotional, behavioral, and cognitive dimensions (Fredricks et al. 2004). The complicated behaviors cause the automatic analysis of engagement in MOOC environment settings to be challenging (Niu et al. 2018). There is no teacher in the MOOC scenario to keep the learners motivated. Thus, there is a requirement to assess

MOOC learner's learning engagement automatically to improve the quality of the learning (Liao et al. 2021; Shen et al. 2021).

Educational institutions are offering online classes to lessen the impact of the covid-19 pandemic. As a result, it is critical to make digital learning sessions engaging and ensure that the students are engaged during the online classes (Bhardwaj et al. 2021). Measuring the engagement level of students automatically, providing real-time feedback, and taking the necessary actions required to achieve the objective is essential in online learning (Abedi and Khan 2021a). The learners in online learning may adapt to any comfortable postures compared to the classroom environment. Thus, the facial expression is one of the most relevant factors to learn engagement that could be easily captured by a web camera during online learning (Shen et al. 2021). So, this work concentrates on learning engagement analysis based on facial expressions rather than body postures.

Instructors encounter difficulty getting timely feedback on each student's engagement and improving the teaching content accordingly. Thus, there is a requirement for automatic prediction of students engagement to enhance the quality of e-learning (Liao et al. 2021). The usage of facial features has shown considerable advantages in the prediction of engagement (Liao et al. 2021). There exist three categories to measure student engagement (i) Self-reports, (ii) Observational checklists and rating scales, and (iii) Automated measurements. The third category, automated measurement, is utilized by researchers in the online learning environment rather than the other two categories due to a lack of temporal resolution (Liao et al. 2021). Automated measurement via computer vision methods is unobtrusive and very well utilized in the past literature (Liao et al. 2021).

In the proposed work, the OpenFace toolkit features and Convolutional Recurrent Neural Network (CRNN) based extracted features are fed into a proposed Facial Engagement Analysis-Network (FEA-Net) for efficient classification of engagement levels. The fine-grained features extracted from OpenFace and CRNN are fed to the deep learning model FEA-Net for further prediction. The main novelty of this work is that within CRNN architecture, Bi-directional Gated Recurrent Unit (Bi-GRU) is

employed, and their intermediate features are extracted and fed into FEA-Net for classification. In FEA-Net, Depthwise separable 1D convolutional layer is used instead of the standard convolution layer. The motivation behind choosing Depthwise separable 1D convolutional layer, it utilizes fewer parameters and less computation. Thus, training the model is faster with less computation power and memory. Center loss (Wen et al. 2016) is employed within the FEA-Net to have discriminative features for engagement recognition.

The contributions of the chapter are :

- A lightweight deep learning FEA-Net architecture is proposed to classify engagement levels.
- The combination of center loss and cross-entropy is introduced with FEA-Net, which helped to improve the discriminative ability of learned features.

The chapter is organized as follows: Section 6.1 is for introduction. Section 6.2 describes the proposed a model for classification of engagement levels. Section 6.3 presents the results and its detailed analysis using various evaluation parameters, and the chapter is concluded in section 6.4.

## 6.2 PROPOSED MODEL FOR FACIAL ENGAGEMENT ANALYSIS

This section gives the overall description of the proposed methodology followed in this work. The overall pipeline of the architecture is shown in Figure 6.1. Initially, sampled frames are retrieved from video data at a predetermined time interval. In the initial stage, face detection is performed using MTCNN. In the next step, the OpenFace toolkit is used for the extraction of handcrafted features. The features like facial landmarks, head poses, eye gaze estimation, and facial Action Units (AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, AU28, AU45) are extracted. Additionally, the intermediate features are extracted using CRNN for the sequence of images using n layers, where n=4. The CRNN model comprises convolution, Batch Normalization, and max-pooling layers.

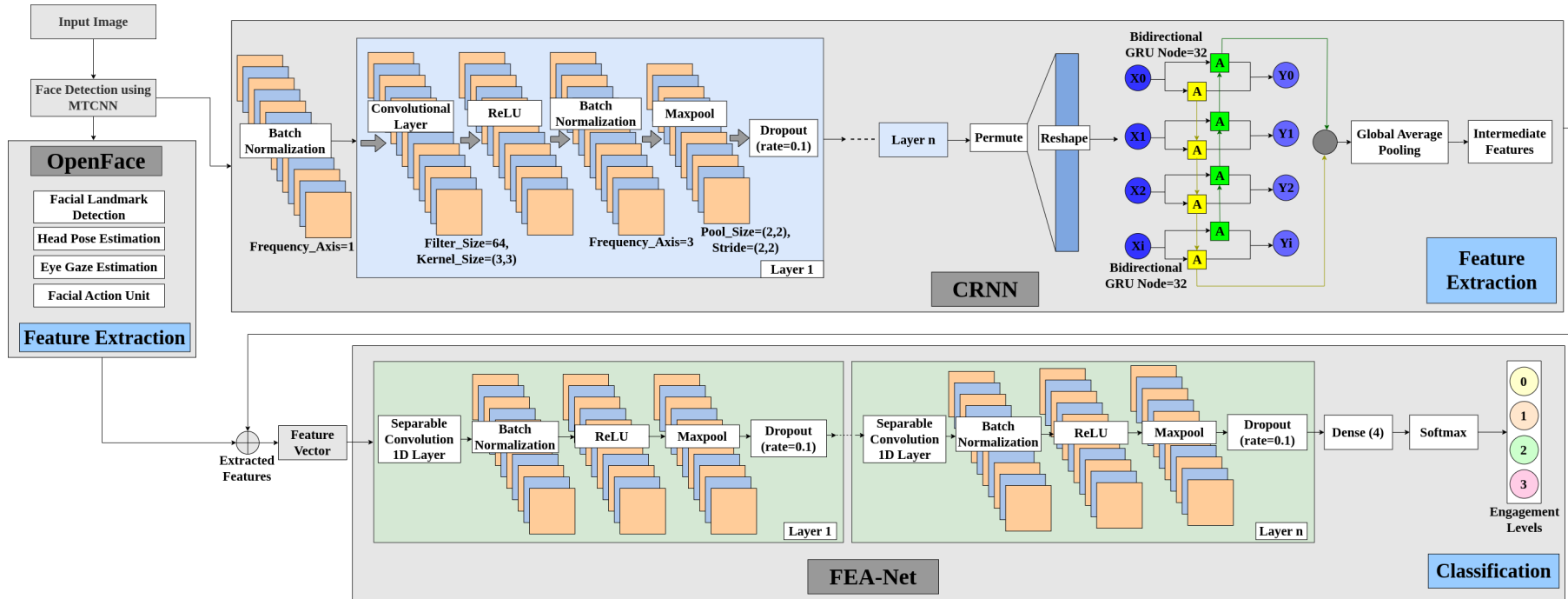


Figure 6.1: Proposed Framework for Engagement Recognition.

Further, the input is reshaped for the recurrent layer and processed to Bi-GRU and global average pooling layers. The intermediate features extracted from the CRNN model are appended with extracted OpenFace features and processed to FEA-Net for classification, where the FEA-Net consists of a depthwise separable convolution layer, Batch Normalization, and ReLU activation function. A loss function with a discriminative ability is proposed, combining center loss and cross-entropy loss within an FEA-Net. Thus, a lightweight model is proposed for the classification of engagement levels.

### 6.2.1 Pre-processing

MTCNN detector performs face detection and facial landmarking jointly, mainly used for face recognition (Ku and Dong 2020). MTCNN (Zhang et al. 2016) is employed to recognize the face region because of its precision and speed. The bounding box is drawn around the faces displaying eyes, nose, and mouth region, and unnecessary boundaries are pruned using MTCNN. The input image is extracted from the video frames. Once the face detection is done, every frame is resized into  $224 \times 224$  input size.

### 6.2.2 Feature Extraction

The usage of facial features in engagement prediction has many advantages, as the face contains many emotional information (Liao et al. 2021). This work employs OpenFace and CRNN features for engagement classification. The elaboration of these features is given in subsections 6.2.2.1 and 6.2.2.2.

#### 6.2.2.1 OpenFace Features

OpenFace 2.0 toolkit is proposed by Baltrusaitis et al. (2018); it has demonstrated exceptional performance on multiple applications such as AUs, head-pose, and eye-gaze identification. OpenFace 2.0 toolkit (Baltrusaitis et al. 2018) is conceived with an intention for researchers working in the computer vision and affective computing community and users who are interested in creating interactive applications using facial behavior analysis. The toolkit can accurately detect facial landmarks, estimate head pose and eye gaze, and facial AU. The behavioral features are extracted

from the OpenFace toolkit according to frame level. The OpenFace toolkit aids in the extraction of high-level facial information from each frame, such as AUs, eye gaze, and head pose aspects. The OpenFace toolkit is open source and does not require numerous human resources. However, reliable facial identification is difficult when the images have low resolution and side profiles (Fydanaki and Geradts 2018).

### 6.2.2.2 Convolutional Recurrent Neural Network (CRNN) Features

The CRNN combines the properties of the Convolutional Neural Network (CNN) and the Recurrent Neural Network (RNN) (Hu et al. 2019; Kim et al. 2020). CNN is used for image classification due to its local receptive fields and weight-sharing nature. In contrast, RNN processes sequential data and takes temporal dimensions into account. In CRNN architecture, CNN is used to extract spatial features from each frame, and their input is reshaped and fed into the BiGRU (Li et al. 2020; Wickramaratne and Mahmud 2020) to extract temporal information, which can aid in the classification of engagement levels.

Gated Recurrent Unit (GRU) is another variant of RNN and is considered as a simpler LSTM with less computational burden (Wickramaratne and Mahmud 2020). GRU gives equal or better performance than the LSTM. RNN using GRU as a hidden unit is considered a powerful and effective model for learning from sequence information. In contrast to LSTM, GRU has gating units that modify information flow within the unit without demanding separate memory cells. Also, in GRU, the number of gate units is lower than in LSTM, and the activation of gate units relies on the current input, and earlier output (Wickramaratne and Mahmud 2020). The model using the GRU hidden unit converges faster due to the parameter reduction, and the ultimate result is preferable over LSTM. BiGRU is a sequential model for extracting characteristics in both directions. Since BiGRU collects data from previous and current steps, it allows for a more thorough examination of the temporal correlation of engagement levels, which in turn improves engagement level classification performance. The proposed CRNN model retrieves information from the image in both forward and backward sequences by applying the BiGRU layer at the output of

the convolutional layer. The CRNN model's intermediate features are loaded, and these features are chosen based on best-saved weights. Further, these features are fed into FEA-Net for classification.

### 6.2.3 Classification using Facial Engagement Analysis-Network (FEA-Net)

The spatial and temporal features generated by CRNN and OpenFace features are fused into FEA-Net, which will help discern the engaged state and improve the performance of the engagement classification. In the FEA-Net, Depthwise separable layer proposed by Sifre and Mallat (2014) is used as a convolutional layer. The Depthwise separable layer has been used in Xception (Extreme Inception) (Chollet 2017) and various other classification applications. This work has utilized light-CNN consisting of four Depthwise separable 1D convolutional layers in FEA-Net for engagement classification.

The Depthwise separable layer is a factorized convolution that divides the standard convolution into two operations such as depthwise convolution, and pointwise convolution that is independent of each other (Ding et al. 2019; Jangra et al. 2021; Khan and Niu 2021). The spatial convolutions are applied independently on each channel in the first phase of depthwise convolution, as shown in Figure 6.2, followed by pointwise convolution, i.e., inter-channel convolution (Le et al. 2021). The primary distinction between a regular convolution layer and a Depthwise separable convolution layer is that each channel of the input layer is convolved using a separate kernel. As a result, lightweight filtering can be done on a single input channel. Next,  $1 \times 1$  pointwise convolution is applied to fuse these input channels linearly. The network parameter is lowered when using Depthwise separable layers instead of a traditional convolution layer, resulting in reduced computational complexity and model size (Ding et al. 2019; Zhang et al. 2019).

## 6.3 RESULTS AND DISCUSSIONS

### 6.3.1 Database Description

The Dataset for Affective States in E-learning Environments (DAiSEE) dataset was utilized to carry out the experiments. The dataset contains four affective states: bored,

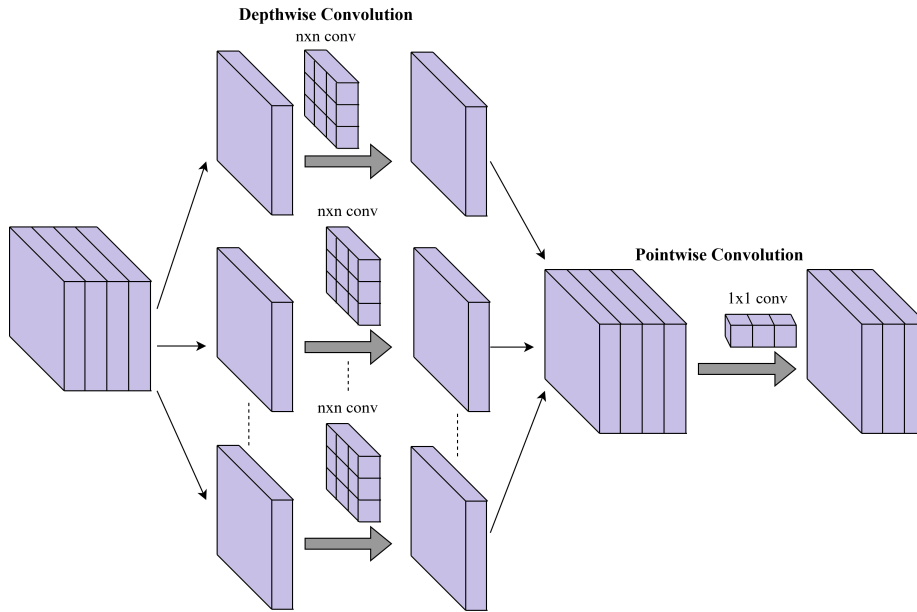


Figure 6.2: Depthwise Separable Convolutional Network architecture (Le et al. 2021)

frustrated, confused, and engaged. Only engagement states are considered for evaluation in this experiment. The engagement level ranges from 0 to 3 (Gupta et al. 2016b; Whitehill et al. 2014), with 0 being very low, 1 being low, 2 being high, and 3 being very high. Table 6.1 shows the sample distribution based on train, test, and validation set on four levels of classes. This distribution is employed with the proposed methodology to have a fair comparison with previous work.

Table 6.1: Dataset Split for Evaluation (Abedi and Khan 2021a)

Level	Train	Validation	Test
0	34	23	4
1	213	143	84
2	2617	813	882
3	2494	450	814

### 6.3.2 Experiment Setup and Neural Network Configuration

The experiments are implemented on an NVIDIA<sup>®</sup> DGX-1<sup>®</sup> system loaded with Canonical Ubuntu Operating System (OS) and an 8X NVIDIA<sup>®</sup> Tesla<sup>®</sup> V100 GPU with 32GB of dedicated RAM. Python libraries such as Keras and Tensorflow (Gulli and Pal 2017) are utilized for the implementation of the proposed model. The CRNN model used for feature extraction takes the input of size 224x224, then the network is



fine-tuned using pre-trained VGGFace (Cao et al. 2018). The extraction of spatial features in the CRNN model is done using layer n, where n=4. Each network layer employs convolutional layers with variable filter sizes. The filter size of 64 is used at the first network layer, the second layer filter size of 128, the third layer filter size of 128, and the final layer filter size of 256 is used. Then the temporal features are extracted from the Bi-GRU layer. Model checkpoint callback is used to store the CRNN model with the best weights, and their intermediate characteristics are fed into the FEA-Net. The dropout layer used in CRNN and FEA-Net avoids overfitting and improves the model's generalization ability. Also, the batch normalization used with these models helps to overcome the further overfitting of the model.

In this training phase, the Adam optimizer is used to update the weights. The learning rate starts at  $1e^{-4}$  with the batch size of 10 and a number of epochs of 50. The early stopping callback is used to stop overlearning of the model during training. The ReLU (Nair and Hinton 2010) is used as activation function to reduce the computations and vanishing gradient problem (Liao et al. 2021). At the final layer of FEA-Net, the softmax activation function is applied. As shown in Equation 6.3, the cross-entropy and center loss is utilized as a final loss function of the model. Due to its nature of acquiring discriminative features (Liao et al. 2021), this loss function is employed within the proposed FEA-Net model.

$$L_{cross-entropy} = - \sum_{i=1}^S \log \left( \frac{\exp(w_{yi}^T \bar{h}_i + b_{yi})}{\sum_{j=1}^C \exp(w_j^T \bar{h}_i + b_j)} \right) \quad (6.1)$$

where  $W=(w_1, w_2 \dots w_C) \in \mathbb{R}^{N \times C}$  and  $b=(b_1, b_2 \dots b_C) \in \mathbb{R}^{1 \times C}$  are the weights and the biases, and S is the batch size and C is the number of categories. The  $\bar{h}_i$  is the attentional hidden layer created by combining the original and new context vectors using concatenation layer (Liao et al. 2021).

$$L_{center} = \frac{1}{2} \sum_{i=1}^S \|\bar{h}_i - k_{yi}\|_2^2 \quad (6.2)$$

where  $k_{yi}$  is the center for the deep features of the  $y_i^{th}$  class.

The total loss function is as follows:

$$L = L_{cross-entropy} + \lambda L_{center} \quad (6.3)$$

The  $\lambda$  coefficient is used to balance the loss.

### 6.3.3 Comparison with state-of-the-art methods

The performance of the proposed model is compared to previous research. As shown in Table 6.2 the proposed model achieves better accuracy than (Abedi and Khan 2021b; Geng et al. 2019; Gupta et al. 2016a; Hao et al. 2019; Liao et al. 2021) but falls short of (Abedi and Khan 2021a) by 1.14% for four-class classification. The TCN has a much longer memory and is better suited to domains requiring a long history (Chen et al. 2021b). Even though Bi-LSTM processes the sequences in both directions, it has lesser efficiency in processing longer history sequences than TCN. The authors (Abedi and Khan (2021a) evaluated the features that are relevant by ranking them using random forest. The proposed methodology, on the other hand, evaluated all features. However, the current work employs a lightweight network, which reduces the model size and allows for faster training.

Table 6.2: Comparison of the proposed model with previous approaches.

Reference	Classifiers used	Accuracy	F1-score
Abedi and Khan (2021a)	TCN	63.30%	NA
Abedi and Khan (2021b)	ResNet +LSTM	61.15%	NA
Abedi and Khan (2021b)	ResNet + TCN	63.90%	NA
Liao et al. (2021)	DFSTN	58.84%	NA
Hao et al. (2019)	I3D	52.35%	NA
Gupta et al. (2016a)	C3D with fine tuning	56.10%	NA
Gupta et al. (2016a)	LRCN	57.90%	NA
Geng et al. (2019)	C3D with focal loss	56.20%	NA
<b>Proposed Work</b>	<b>OpenFace + CRNN → FEA-Net</b>	<b>62.16%</b>	<b>0.3954</b>

\* NA–Not Available

#### 6.3.3.1 Discussions

There is an extreme imbalance in the four levels of engagement, which will lower the prediction rate. Very low engagement and low engagement samples are two groups that are causing some confusion. Further, more importance needs to be given to these two classes (engagement and low engagement) for proper classification. Labeling two classes very high engagement and high engagement classes appropriately is also a challenging task, thus reducing the system’s performance. Figure 6.3 presents a sample of images depicting the four levels of engagement.

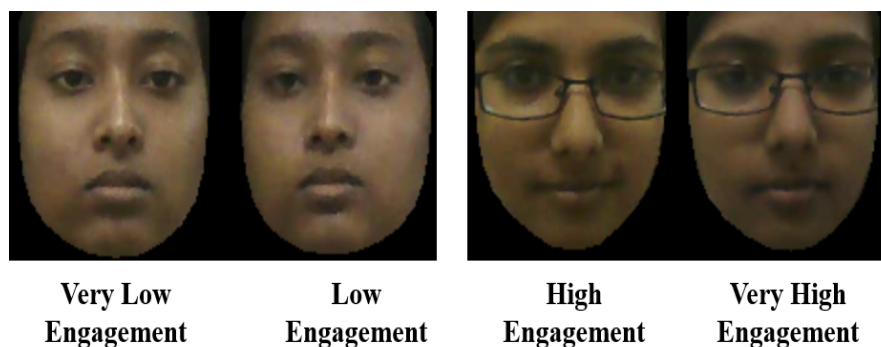


Figure 6.3: Four levels of engagement from DAiSEE Dataset

The key benefit of adopting FEA-Net is that it uses depthwise separable convolution, which uses fewer parameters than regular convolutional layers and is less susceptible to overfitting. The model computes with fewer operations, making it cheaper and faster to process. The OpenFace features such as landmark points and their motion, head pose (orientation and movement), eye gaze, and AUs; all these modalities individually and together play an essential role in analyzing human behavior. These features help FEA-Net capture valuable features for the classification of engagement levels.

#### 6.4 SUMMARY

In this chapter, a novel FEA-Net architecture is proposed to classify student's engagement levels. The OpenFace and CRNN features are combined in the FEA-Net, and intraclass variances are reduced using center loss and cross-entropy. The Depthwise separable convolution layer employed with FEA-Net reduces the model complexity. The combination of spatial and temporal features, OpenFace features fed into FEA-Net for classification obtained an accuracy of 62.16%. Future work would be to reduce the misclassified labels and re-train the images of those classes that were predicted wrong, thus improving the classification performance. Also, resolve the issue of a significant data imbalance. Further, the model's classification accuracy could be increased using the deep layers in FEA-Net.



## **CHAPTER 7**

### **CONCLUSIONS AND FUTURE SCOPE**

This chapter concludes the work, along with some potential areas for additional research. The development of the FER system depends on a number of subtasks, including occlusion recovery, partial face recognition, occlusion discard face recognition and facial expression recognition, personalized emotion recognition, gender-based facial expression recognition, and others; each system's implementation has a positive impact on the development of the FER system. This thesis considers a few essential tasks, addressing pose variations and occlusions which are some of the computer vision challenges, and also classifies the posed and non-posed expressions using machine and deep learning approaches. This chapter summarizes the work addressed in the thesis, provides learning outcomes as conclusions, and emphasizes a few issues as potential areas for further research.

#### **7.1 SUMMARY AND CONCLUSIONS**

Facial expression plays a vital role in interpersonal communication. Automatic analysis of facial expressions is crucial to Human-Computer Interaction (HCI) and has become an interesting area over the past decades. The thesis's first work focused on detecting non-posed expressions. Micro-Expressions (ME) are non-posed (spontaneous) expressions with facial muscle movements that last for not more than 200ms. MEs occur in distinct facial regions; hence, they combine distinct facial regions instead of considering only one facial area. This motivated us to select the

most significant facial areas based on AU indexes. The work focused on AU analysis and considered the contribution of each AU to the analysis of MEs, as each AU contributes differently to the classification of facial expressions. Distinct AUs are modeled, and features are extracted and integrated to help in the identification of minute changes over the face. Since it is tedious and challenging to generate facial representations for vast quantities of AUs, this work focused on finding the representations for AUs of only important facial regions that would strongly help in the classification of MEs. The proposed MER model helped distinguish minute changes in facial areas and efficiently classify MEs with accuracies of 76.47% and 67.19% when performing the cross-dataset evaluation using ME databases. The extraction of relevant features from the important facial region based on AU indexes greatly enhanced the ME-related features and aided the recognition of MEs. Thus, the proposed MER model helps in ME identification and overcoming challenges like subtle variations in a short time and intricate interplay between facial areas.

The next work focused on detecting facial occlusion, which frequently occurs in natural settings and is challenging in computer vision and object detection. The work focused on detecting a subset of typical occlusions caused by external accessories. The proposed work performs simultaneous categorization of an occluded and non-occluded face and identifies the occluded faces efficiently. The difficulty of obscuring faces was devastated by the localization of the facial occlusions using the Xcep-RA network, which showed a performance of 99.85% and 98.95% on occluded datasets. The performance of occlusion detection further aids in the effective selection of features and lead to more accurate recognition by minimizing data loss.

Further, the thesis work focused on detecting posed expressions using deep neural networks. To detect posed expressions, selecting appropriate features to distinguish individuals' emotions from various categories of emotions is essential. Thus, selecting efficient feature extraction and classification techniques is the key challenge in recognizing facial expressions efficiently. The multi-stage classification approach has been proposed in this work. At the initial stage, distinct features are extracted from the EfficientNets intermediate layer to categorize posed expressions. Further, two

ensemble models are proposed to classify facial expressions into respective expression classes. The proposed multi-stage posed expression classification model gave the best results with the majority voting and stacking classifier approach; and achieved accuracies of 98.71% and 98.56%, respectively, on posed datasets, making the proposed system robust against pose variations. It is observed that combining multiple classifiers induces higher-level classifiers (meta classifiers) and tries to learn all possible patterns (learns from the errors) from the base classifiers before making the final prediction. Thus, by integrating the outputs from the base classifiers, the ensemble model makes accurate predictions, reduces overfitting, and reduces the risk of getting varied outputs from different machine learning classifiers.

The final work of the thesis focused on developing a system for the analysis of engagement levels. Engagement recognition is essential for monitoring online learning for efficient learning outcomes. By monitoring the student's engagement, the teacher will acquire timely feedback, diminish the dropout rates, and overcome educational problems. A Facial Engagement Analysis-Network (FEA-Net) is proposed for learning engagement assessment in Massive Open Online Courses (MOOC) scenarios. In a MOOC setting, the combination of CRNN and OpenFace features fed into FEA-Net proved effective for classifying engagement levels. The proposed FEA-Net built using the Depthwise Separable Convolution layer helped improve the system's performance by reducing the model complexity. The model achieved an accuracy of 62.16% on the Dataset for Affective States in E-learning Environments (DAiSEE) dataset. It is observed that using a depthwise separable convolution layer within FEA-Net reduced the overfitting of the model. The OpenFace features integrated with CRNN features played an essential role in analyzing human behavior, which helped FEA-Net capture valuable features that would aid in classifying engagement levels.

Finally, the inference obtained from the current research is that ensemble models with the help of multi-stage classification help to reduce the errors from propagating to higher-level classifiers. This method can be successfully utilized to carry out other FER tasks. Further, identifying predominant AUs and extracting features from those dominant regions are essential in identifying minute changes that occur over the face;

Thus, it is helpful in high-risk situations and various real-time applications. Also, identifying the occluded areas and reducing the difficulty of obscurity would help identify the faces more prominently.

## **7.2 LIMITATIONS AND POSSIBLE FUTURE DIRECTIONS**

The work presented in this thesis may be further extended and improved as follows

1. In the current research, the facial region is segmented using Delaunay Triangulation (DT) and Voronoi Diagram (VD) techniques; in the future, more effective segmentation methodologies can be taken into consideration to segment significant facial regions. Additionally, the features can be automatically retrieved from those regions utilizing CNN architectures. This extension might produce an automated FER system appropriate for use in practical applications.
2. The mapping of facial deformations into AUs and adequate standardization of such deformations are necessary. Additionally, a strong correlation between the mapping of AUs and expressions needs to be developed in the near future.
3. Effective analysis of facial expressions heavily depends on accurate representation of facial features. The features that are computed for expression recognition have been directly utilized to perform FER tasks without thorough analysis. Hence, a standard correlation analysis is essential in deciding the feature set to help reduce dimensionality and minimize computational complexity.
4. The current research has performed the localization of the occluded areas. Further, this work can be extended by repairing the occluded images using reconstruction and inpainting approaches and performing face and facial expression recognition tasks.
5. Since the occluded area of the face image might vary in position, size, and form. Gathering a large-scale training dataset containing every form and distinct type of occlusion is essential to train the model and thus overcome the challenges of the realistic scenario.



6. The multi-stage posed classification model is proposed in the current research. This work can be extended to find the similarities of features within the class and maximize the variations between classes. By doing so, the misclassification or false predictions between the classes, like disgust and fear, fear and sadness, can be reduced. Also, more weightage can be given to important facial regions that help to reduce the false predictions between such categories of expressions.
7. In the current work, only facial engagement recognition to assess the engagement levels of students in MOOC scenarios is considered. This work can be extended by considering other affective dimensions, such as bored, frustration, confusion, fear, etc., in online learning.
8. Further, the work can be extended in evaluating multiple faces, for example, the student's engagement in a classroom environment, by considering the emotional and behavioral features which commonly occur in natural settings.
9. A personalized and generalized system that adapts to time and circumstance has not been the focus of the current work. This work can be extended further, considering these aspects into consideration.



## REFERENCES

- Abate, A. F., Cimmino, L., Mocanu, B.-C., Narducci, F. and Pop, F. (2023). “The limitations for expression recognition in computer vision introduced by facial masks.” *Multimedia Tools and Applications*, 82(8), 11305–11319.
- Abedi, A. and Khan, S. (2021a). “Affect-driven engagement measurement from videos.” *arXiv preprint arXiv:2106.10882*.
- Abedi, A. and Khan, S. S. (2021b). “Improving state-of-the-art in detecting student engagement with resnet and tcn hybrid network.” *arXiv preprint arXiv:2104.10122*.
- Adolphs, R. (2002). “Recognizing emotion from facial expressions: psychological and neurological mechanisms.” *Behavioral and cognitive neuroscience reviews*, 1(1), 21–62.
- Aggarwal, C. C. (2015). “Data classification.” In *Data mining*, Springer, 285–344.
- Agrawal, A. and Mittal, N. (2020). “Using cnn for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy.” *The Visual Computer*, 36(2), 405–412.
- Álvarez, A., Sierra, B., Arruti, A., López-Gil, J.-M. and Garay-Vitoria, N. (2016). “Classifier subset selection for the stacked generalization method applied to emotion recognition in speech.” *Sensors*, 16(1), 21–46.
- Azmi, R. and Yegane, S. (2012). “Facial expression recognition in the presence of occlusion using local gabor binary patterns.” In *20th Iranian Conference on Electrical Engineering (ICEE2012)*, IEEE, 742–747.
- Baltrusaitis, T., Zadeh, A., Lim, Y. C. and Morency, L.-P. (2018). “Openface 2.0: Facial behavior analysis toolkit.” In *13th International Conference on Automatic Face & Gesture Recognition (FG 2018)*, IEEE, 59–66.
- Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M. and Pollak, S. D. (2019). “Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements.” *Psychological science in the public interest*, 20(1), 1–68.
- Barros, P., Parisi, G. I., Weber, C. and Wermter, S. (2017). “Emotion-modulated attention improves expression recognition: A deep learning model.” *Neurocomputing*, 253, 104–114.
- Bebis, G., Deaconu, T. and Georgiopoulos, M. (1999). “Fingerprint identification using delaunay triangulation.” In *International Conference on Information Intelligence and Systems*, IEEE, 452–459.
- Bhardwaj, P., Gupta, P., Panwar, H., Siddiqui, M. K., Morales-Menendez, R. and Bhaik, A. (2021). “Application of deep learning on student engagement in e-learning environments.” *Computers & Electrical Engineering*, 93, 107277–107287.

## REFERENCES

---

- Bhushan, B. (2015). “Study of facial micro-expressions in psychology.” In *Understanding facial expressions in communication*, Springer, 265–286.
- Boughida, A., Kouahla, M. N. and Laffi, Y. (2022). “A novel approach for facial expression recognition based on gabor filters and genetic algorithm.” *Evolving Systems*, 13(2), 331–345.
- Cai, J., Meng, Z., Khan, A. S., O’Reilly, J., Li, Z., Han, S. and Tong, Y. (2021). “Identity-free facial expression recognition using conditional generative adversarial network.” In *IEEE International Conference on Image Processing (ICIP)*, IEEE, 1344–1348.
- Cao, Q., Shen, L., Xie, W., Parkhi, O. M. and Zisserman, A. (2018). “Vggface2: A dataset for recognising faces across pose and age.” In *13th International Conference on Automatic Face & Gesture Recognition (FG 2018)*, IEEE, 67–74.
- Carcagnì, P., Del Coco, M., Leo, M. and Distantè, C. (2015). “Facial expression recognition and histograms of oriented gradients: a comprehensive study.” *SpringerPlus*, 4(1), 1–25.
- Cheddad, A., Mohamad, D. and Abd Manaf, A. (2008). “Exploiting voronoi diagram properties in face segmentation and feature extraction.” *Pattern Recognition*, 41(12), 3842–3859.
- Chen, B., Guan, W., Li, P., Ikeda, N., Hirasawa, K. and Lu, H. (2021a). “Residual multi-task learning for facial landmark localization and expression recognition.” *Pattern Recognition*, 115, 107893–107901.
- Chen, J., Chen, D. and Liu, G. (2021b). “Using temporal convolution network for remaining useful lifetime prediction.” *Engineering Reports*, 3(3), Art. no. e12305.
- Chen, J., Chen, Z., Chi, Z. and Fu, H. (2016). “Facial expression recognition in video with multiple feature fusion.” *IEEE Transactions on Affective Computing*, 9(1), 38–50.
- Chen, Y., Song, L., Hu, Y. and He, R. (2018). “Adversarial occlusion-aware face detection.” In *9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, IEEE, 1–9.
- Chollet, F. (2017). “Xception: Deep learning with depthwise separable convolutions.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 1251–1258.
- Cohn, J. F., Ambadar, Z. and Ekman, P. (2007). “Observer-based measurement of facial expression with the facial action coding system.” *The handbook of emotion elicitation and assessment*, 203–221.
- Cruz, A. C., Bhanu, B. and Thakoor, N. S. (2014). “Vision and attention theory based sampling for continuous facial emotion recognition.” *IEEE Transactions on Affective Computing*, 5(4), 418–431.
- Cugu, I., Sener, E. and Akbas, E. (2019). “Microexpnet: An extremely small and fast model for expression recognition from face images.” In *9th International Conference on Image Processing Theory, Tools and Applications (IPTA)*, IEEE, 1–6.
- Dagnes, N., Vezzetti, E., Marcolin, F. and Tornincasa, S. (2018). “Occlusion detection and restoration techniques for 3d face recognition: a literature review.” *Machine Vision and Applications*, 29, 789–813.

- Dailey, M. N., Joyce, C., Lyons, M. J., Kamachi, M., Ishi, H., Gyoba, J. and Cottrell, G. W. (2010). "Evidence and a computational explanation of cultural differences in facial expression recognition." *Emotion*, 10(6), 874–893.
- Dandan, L., Liang, H., Yu, Z. and Zhang, Y. (2020). "Deep convolutional bilstm fusion network for facial expression recognition." *The Visual Computer*, 36(3), 499–508.
- Darwin, C. and Prodger, P. (1998). *The expression of the emotions in man and animals*, Oxford University Press, USA.
- Davison, A. K., Lansley, C., Costen, N., Tan, K. and Yap, M. H. (2016). "Samm: A spontaneous micro-facial movement dataset." *IEEE Transactions on Affective Computing*, 9(1), 116–129.
- De la Torre, F. and Cohn, J. F. (2011). "Facial expression analysis." *Visual analysis of humans: Looking at people*, 377–409.
- Dhall, A., Goecke, R., Joshi, J., Sikka, K. and Gedeon, T. (2014). "Emotion recognition in the wild challenge 2014: Baseline, data and protocol." In *16th International Conference on Multimodal Interaction*, ACM, 461–466.
- Dhall, A., Goecke, R., Lucey, S. and Gedeon, T. (2011). "Acted facial expressions in the wild database." *Australian National University, Canberra, Australia, Technical Report TR-CS-11*, 2, 1–15.
- Din, N. U., Javed, K., Bae, S. and Yi, J. (2020). "A novel gan-based network for unmasking of masked face." *IEEE Access*, 8, 44276–44287.
- Ding, H., Zhou, S. K. and Chellappa, R. (2017). "Facenet2expnet: Regularizing a deep face recognition net for expression recognition." In *12th International Conference on Automatic Face & Gesture Recognition (FG 2017)*, IEEE, 118–126.
- Ding, W., Huang, Z., Huang, Z., Tian, L., Wang, H. and Feng, S. (2019). "Designing efficient accelerator of depthwise separable convolutional neural network on fpga." *Journal of Systems Architecture*, 97, 278–286.
- Dino, H. I. and Abdulrazzaq, M. B. (2019). "Facial expression classification based on svm, knn and mlp classifiers." In *International Conference on Advanced Science and Engineering (ICOASE)*, IEEE, 70–75.
- Dodia, S., Annappa, B. and Padukudru, M. A. (2022). "A novel bi-level lung cancer classification system on ct scans." In *26th Annual Conference on Medical Image Understanding and Analysis, MIUA 2022*, Springer, 578–593.
- Dong, Z., Wang, G., Lu, S., Li, J., Yan, W. and Wang, S.-J. (2022). "Spontaneous facial expressions and micro-expressions coding: From brain to face." *Frontiers in Psychology*, Art. No. 784834.
- Du, S. and Martinez, A. M. (2022). "Compound facial expressions of emotion: from basic research to clinical applications." *Dialogues in clinical neuroscience*, 17(4), 443–455.
- Du, S., Tao, Y. and Martinez, A. M. (2014). "Compound facial expressions of emotion." *Proceedings of the National Academy of Sciences*, 111(15), E1454–E1462.
- Edla, D. R., Ansari, M. F., Chaudhary, N. and Dodia, S. (2018). "Classification of facial expressions from eeg signals using wpt and svm for wheelchair control operations." *Procedia Computer Science*, 132, 1467–1476.

## REFERENCES

---

- Ekenel, H. K. and Stiefelhagen, R. (2009). “Why is facial occlusion a challenging problem?” In *International Conference on Biometrics*, Springer, 299–308.
- Ekman, P. (1964). “Body position, facial expression, and verbal behavior during interviews.” *The Journal of Abnormal and Social Psychology*, 68(3), 295–301.
- Ekman, P. (2009a). “Darwin’s contributions to our understanding of emotional expressions.” *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535), 3449–3451.
- Ekman, P. (2009b). “Lie catching and microexpressions.” *The philosophy of deception*, 118–133.
- Ekman, P. and Friesen, W. V. (1969). “Nonverbal leakage and clues to deception.” *Psychiatry*, 32(1), 88–106.
- Ekman, P. and Friesen, W. V. (1971). “Constants across cultures in the face and emotion.” *Journal of personality and social psychology*, 17(2), 124–129.
- Ekman, P. and Friesen, W. V. (1978). *Facial action coding system: Investigator’s guide*, Consulting Psychologists Press.
- Fan, Y., Lam, J. C. and Li, V. O. (2018). “Multi-region ensemble convolutional neural network for facial expression recognition.” In *International Conference on Artificial Neural Networks*, Springer, 84–94.
- Farzaneh, A. H. and Qi, X. (2021). “Facial expression recognition in the wild via deep attentive center loss.” In *Winter Conference on Applications of Computer Vision*, IEEE, 2402–2411.
- Fathallah, A., Abdi, L. and Douik, A. (2017). “Facial expression recognition via deep learning.” In *14th International Conference on Computer Systems and Applications (AICCSA)*, IEEE, 745–750.
- Feifei, Z., Yu, Y., Mao, Q., Gou, J. and Zhan, Y. (2016). “Pose-robust feature learning for facial expression recognition.” *Frontiers of Computer Science*, 10(5), 832–844.
- Feifei, Z., Zhang, T., Mao, Q. and Xu, C. (2018). “Joint pose and expression modeling for facial expression recognition.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 3359–3368.
- Fredricks, J. A., Blumenfeld, P. C. and Paris, A. H. (2004). “School engagement: Potential of the concept, state of the evidence.” *Review of educational research*, 74(1), 59–109.
- Fydanaki, A. and Geradts, Z. (2018). “Evaluating openface: an open-source automatic facial comparison algorithm for forensics.” *Forensic sciences research*, 3(3), 202–209.
- Ge, H., Zhu, Z., Dai, Y., Wang, B. and Wu, X. (2022). “Facial expression recognition based on deep learning.” *Computer Methods and Programs in Biomedicine*, 106621–106629.
- Geng, L., Xu, M., Wei, Z. and Zhou, X. (2019). “Learning deep spatiotemporal feature for engagement recognition of online courses.” In *Symposium Series on Computational Intelligence (SSCI)*, IEEE, 442–447.
- Georgescu, M.-I., Ionescu, R. T. and Popescu, M. (2019). “Local learning with deep and handcrafted features for facial expression recognition.” *IEEE Access*, 7, 64827–64836.

- Ghazouani, H. (2021). “A genetic programming-based feature selection and fusion for facial expression recognition.” *Applied Soft Computing*, 103, 107173–107186.
- Ghiassi, G., Fowlkes, C. C. and Irvine, C. (2015). “Using segmentation to predict the absence of occluded parts.” In *BMVC*, Citeseer, 22–1.
- Ghimire, D., Jeong, S., Lee, J. and Park, S. H. (2017). “Facial expression recognition based on local region specific features and support vector machines.” *Multimedia Tools and Applications*, 76(6), 7803–7821.
- Ghimire, D. and Lee, J. (2013). “Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines.” *Sensors*, 13(6), 7714–7734.
- Gizatdinova, Y. and Surakka, V. (2008). “Effect of facial expressions on feature-based landmark localization in static grey scale images.” In *International Conference on Computer Vision Theory and Applications (VISAPP)*, IEEE, 259–266.
- González-Hernández, F., Zatarain-Cabada, R., Barrón-Estrada, M. L. and Rodríguez-Rangel, H. (2018). “Recognition of learning-centered emotions using a convolutional neural network.” *Journal of Intelligent & Fuzzy Systems*, 34(5), 3325–3336.
- Gulli, A. and Pal, S. (2017). *Deep learning with Keras*, Packt Publishing Ltd.
- Guo, C., Liang, J., Zhan, G., Liu, Z., Pietikäinen, M. and Liu, L. (2019). “Extended local binary patterns for efficient and robust spontaneous facial micro-expression recognition.” *IEEE Access*, 7, 174517–174530.
- Guo, J., Lei, Z., Wan, J., Avots, E., Hajarolasvadi, N., Knyazev, B., Kuharenko, A., Junior, J. C. S. J., Baró, X., Demirel, H. et al. (2018). “Dominant and complementary emotion recognition from still images of faces.” *IEEE Access*, 6, 26391–26403.
- Gupta, A., D’Cunha, A., Awasthi, K. and Balasubramanian, V. (2016a). “Daisee: Towards user engagement recognition in the wild.” *arXiv preprint arXiv:1609.01885*.
- Gupta, A., Jaiswal, R., Adhikari, S. and Balasubramanian, V. (2016b). “Daisee: dataset for affective states in e-learning environments.” *arXiv preprint arXiv:1609.01885*, 1–22.
- Hao, Z., Xiao, X., Huang, T., Liu, S., Xia, Y. and Li, J. (2019). “An novel end-to-end network for automatic student engagement recognition.” In *9th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, IEEE, 342–345.
- Happy, S., Dantcheva, A. and Bremond, F. (2019). “A weakly supervised learning technique for classifying facial expressions.” *Pattern Recognition Letters*, 128, 162–168.
- Happy, S. and Routray, A. (2014). “Automatic facial expression recognition using features of salient facial patches.” *IEEE Transactions on Affective Computing*, 6(1), 1–12.
- He, J., Yu, X., Sun, B. and Yu, L. (2021). “Facial expression and action unit recognition augmented by their dependencies on graph convolutional networks.” *Journal on Multimodal User Interfaces*, 1–12.
- Hemathilaka, S. and Aponso, A. (2022). “A comprehensive study on occlusion invariant face recognition under face mask occlusion.” *arXiv preprint arXiv:2201.09089*.

## REFERENCES

---

- Hess, U. and Thibault, P. (2009). “Darwin and emotion expression.” *American Psychologist*, 64(2), 120–128.
- Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V. et al. (2019). “Searching for mobilenetv3.” In *International Conference on Computer Vision*, IEEE, 1314–1324.
- Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H. (2017). “Mobilenets: Efficient convolutional neural networks for mobile vision applications.” *arXiv preprint arXiv:1704.04861*.
- Hu, Z., Hu, Y., Liu, J., Wu, B., Han, D. and Kurfess, T. (2019). “A crnn module for hand pose estimation.” *Neurocomputing*, 333, 157–168.
- Huang, B., Wang, Z., Jiang, K., Zou, Q., Tian, X., Lu, T. and Han, Z. (2022). “Joint segmentation and identification feature learning for occlusion face recognition.” *IEEE Transactions on Neural Networks and Learning Systems*, 1–14.
- Huang, B., Wang, Z., Wang, G., Jiang, K., Zeng, K., Han, Z., Tian, X. and Yang, Y. (2021). “When face recognition meets occlusion: A new benchmark.” In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 4240–4244.
- Hung, J. C., Lin, K.-C. and Lai, N.-X. (2019). “Recognizing learning emotion based on convolutional neural networks and transfer learning.” *Applied Soft Computing*, 105724–105742.
- Jangra, M., Dhull, S. K., Singh, K. K., Singh, A. and Cheng, X. (2021). “O-wcnn: an optimized integration of spatial and spectral feature map for arrhythmia classification.” *Complex & Intelligent Systems*, 1–14.
- Jia, X., Ben, X., Yuan, H., Kpalma, K. and Meng, W. (2018). “Macro-to-micro transformation model for micro-expression recognition.” *Journal of computational science*, 25, 289–297.
- Jiang, B. and Jia, K. (2013). “Semi-supervised facial expression recognition algorithm on the condition of multi-pose.” *Journal of Information Hiding and Multimedia Signal Processing*, 4, 138–146.
- Jiang, B. and Jia, K. (2016). “Robust facial expression recognition algorithm based on local metric learning.” *Journal of Electronic Imaging*, 25(1), 013022–013030.
- Jiang, S. and Jiang, W. (2019). “Reliable image matching via photometric and geometric constraints structured by delaunay triangulation.” *ISPRS Journal of Photogrammetry and Remote Sensing*, 153, 1–20.
- Jiang, W., Ye, L., Yi, Z. and Peng, C. (2022). “A new occluded face recognition framework with combination of both deocclusion and feature filtering methods.” *Multimedia Tools and Applications*, 81(23), 33867–33896.
- Jung, H., Lee, S., Yim, J., Park, S. and Kim, J. (2015). “Joint fine-tuning in deep neural networks for facial expression recognition.” In *International Conference on Computer Vision*, IEEE, 2983–2991.
- Kaya, H., Gürpınar, F. and Salah, A. A. (2017). “Video-based emotion recognition in the wild using deep transfer learning and score fusion.” *Image and Vision Computing*, 65, 66–75.



- Khan, Z. Y. and Niu, Z. (2021). “Cnn with depthwise separable convolutions and combined kernels for rating prediction.” *Expert Systems with Applications*, 170, 114528–114536.
- Khodadoust, J. and Khodadoust, A. M. (2017). “Fingerprint indexing based on expanded delaunay triangulation.” *Expert Systems with Applications*, 81, 251–267.
- Khor, H.-Q., See, J., Liong, S.-T., Phan, R. C. and Lin, W. (2019). “Dual-stream shallow networks for facial micro-expression recognition.” In *International Conference on Image Processing (ICIP)*, IEEE, 36–40.
- Khorrami, P., Paine, T. and Huang, T. (2015). “Do deep neural networks learn facial action units when doing expression recognition?.” In *International Conference on Computer Vision Workshops*, IEEE, 19–27.
- Kim, C.-M., Hong, E. J., Chung, K. and Park, R. C. (2020). “Driver facial expression analysis using lfa-crn-based feature extraction for health-risk decisions.” *Applied Sciences*, 10(8), 133–160.
- Kim, D. H., Baddar, W. J., Jang, J. and Ro, Y. M. (2017). “Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition.” *IEEE Transactions on Affective Computing*, 10(2), 223–236.
- Ko, B. (2018). “A brief review of facial emotion recognition based on visual information.” *sensors*, 18(2), 401–420.
- Kotsia, I., Nikolaidis, N. and Pitas, I. (2006). “Fusion of geometrical and texture information for facial expression recognition.” In *International Conference on Image Processing*, IEEE, 2649–2652.
- Koujan, M. R., Alharbawee, L., Giannakakis, G., Pugeault, N. and Roussos, A. (2020). “Real-time facial expression recognition “in the wild” by disentangling 3d expression from identity.” In *15th International Conference on Automatic Face and Gesture Recognition (FG 2020)*, IEEE, 24–31.
- Ku, H. and Dong, W. (2020). “Face recognition based on mtcnn and convolutional neural network.” *Frontiers in Signal Processing*, 4(1), 37–42.
- Kuang, L., Zhang, M. and Pan, Z. (2016). “Facial expression recognition with cnn ensemble.” In *International Conference on Cyberworlds (CW)*, IEEE, 163–166.
- Kumar, A. J. R., Theagarajan, R., Peraza, O. and Bhanu, B. (2019). “Classification of facial micro-expressions using motion magnified emotion avatar images.” In *Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, IEEE, 12–20.
- Kun-Hong, L., Jin, Q.-S., Xu, H.-C., Gan, Y.-S. and Liong, S.-T. (2021). “Micro-expression recognition using advanced genetic algorithm.” *Signal Processing: Image Communication*, 93, 116153–116162.
- Kurup, A. R., Ajith, M. and Ramón, M. M. (2019). “Semi-supervised facial expression recognition using reduced spatial features and deep belief networks.” *Neurocomputing*, 367, 188–197.
- Lahasan, B., Lutfi, S. L. and San-Segundo, R. (2019). “A survey on techniques to handle face recognition challenges: occlusion, single sample per subject and expression.” *Artificial Intelligence Review*, 52, 949–979.

## REFERENCES

---

- Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T. and Van Knippenberg, A. (2010). “Presentation and validation of the radboud faces database.” *Cognition and emotion*, 24(8), 1377–1388.
- Le, D.-N., Parvathy, V. S., Gupta, D., Khanna, A., Rodrigues, J. J. and Shankar, K. (2021). “Tot enabled depthwise separable convolution neural network with deep support vector machine for covid-19 diagnosis and classification.” *International Journal of Machine Learning and Cybernetics*, 12(11), 3235–3248.
- Learned-Miller, E., Huang, G. B., RoyChowdhury, A., Li, H. and Hua, G. (2016). “Labeled faces in the wild: A survey.” *Advances in face detection and facial image analysis*, 189–248.
- Lee, C. H., Liu, Z., Wu, L. and Luo, P. (2020). “Maskgan: Towards diverse and interactive facial image manipulation.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 5549–5558.
- Lee, K., Yoon, H., Song, J. and Park, K. (2018). “Convolutional neural network-based classification of driver’s emotion during aggressive and smooth driving using multi-modal camera sensors.” *Sensors*, 18(4), 957–978.
- Lee, S. H. and Ro, Y. M. (Oct.-Dec. 2016). “Partial matching of facial expression sequence using over-complete transition dictionary for emotion recognition.” *IEEE Transactions on Affective Computing*, 7(4), 389–408.
- Lei, L., Li, J., Chen, T. and Li, S. (2020). “A novel graph-tcn with a graph structured representation for micro-expression recognition.” In *28th International Conference on Multimedia*, ACM, 2237–2245.
- Li, P., Luo, A., Liu, J., Wang, Y., Zhu, J., Deng, Y. and Zhang, J. (2020). “Bidirectional gated recurrent unit neural network for chinese address element segmentation.” *ISPRS International Journal of Geo-Information*, 9(11), 635–653.
- Li, S. and Deng, W. (2018a). “Deep facial expression recognition: A survey.” *arXiv preprint arXiv:1804.08348*.
- Li, S. and Deng, W. (2018b). “Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition.” *IEEE Transactions on Image Processing*, 28(1), 356–370.
- Li, S., Deng, W. and Du, J. (2017). “Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2852–2861.
- Li, W. and Zou, L. (2017). “Classifier stacking for native language identification.” In *12th Workshop on Innovative Use of NLP for Building Educational Applications*, 390–397.
- Li, X., Pfister, T., Huang, X., Zhao, G. and Pietikäinen, M. (2013). “A spontaneous micro-expression database: Inducement, collection and baseline.” In *10th International Conference and Workshops on Automatic face and gesture recognition (fg)*, IEEE, 1–6.
- Li, Y., Huang, X. and Zhao, G. (2021). “Micro-expression action unit detection with spatial and channel attention.” *Neurocomputing*, 436, 221–231.
- Li, Y., Lu, Y., Li, J. and Lu, G. (2019). “Separate loss for basic and compound facial expression recognition in the wild.” In *Asian Conference on Machine Learning (ACML)*, Springer, 897–911.

- Li, Y., Zeng, J., Shan, S. and Chen, X. (2018). “Occlusion aware facial expression recognition using cnn with attention mechanism.” *IEEE Transactions on Image Processing*, 28(5), 2439–2450.
- Liang, L., Lang, C., Li, Y., Feng, S. and Zhao, J. (2020). “Fine-grained facial expression recognition in the wild.” *IEEE Transactions on Information Forensics and Security*, 16, 482–494.
- Liao, J., Liang, Y. and Pan, J. (2021). “Deep facial spatiotemporal network for engagement prediction in online learning.” *Applied Intelligence*, 1–13.
- Liong, S.-T., See, J., Wong, K. and Phan, R. C.-W. (2018). “Less is more: Micro-expression recognition from video using apex frame.” *Signal Processing: Image Communication*, 62, 82–92.
- Liong, S.-T. and Wong, K. (2017). “Micro-expression recognition using apex frame with phase information.” In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 534–537.
- Liu, M., Shan, S., Wang, R. and Chen, X. (2016). “Learning expressionlets via universal manifold model for dynamic facial expression recognition.” *IEEE Transactions on Image Processing*, 25(12), 5920–5932.
- Liu, P., Lin, Y., Meng, Z., Deng, W., Zhou, J. T. and Yang, Y. (2020). “Point adversarial self mining: A simple method for facial expression recognition in the wild.” *arXiv preprint arXiv:2008.11401*.
- Liu, Y., Dai, W., Fang, F., Chen, Y., Huang, R., Wang, R. and Wan, B. (2021). “Dynamic multi-channel metric network for joint pose-aware and identity-invariant facial expression recognition.” *Information Sciences*, 578, 195–213.
- Liu, Y., Yuan, X., Gong, X., Xie, Z., Fang, F. and Luo, Z. (2018). “Conditional convolution neural network enhanced random forest for facial expression recognition.” *Pattern Recognition*, 84, 251–261.
- Liu, Y., Zhang, L., Hao, Z., Yang, Z., Wang, S., Zhou, X. and Chang, Q. (2022). “An exception model based on residual attention mechanism for the classification of benign and malignant gastric ulcers.” *Scientific Reports*, 12(1), Art No. 15365.
- Liu, Y., Zhang, X., Lin, Y. and Wang, H. (2019). “Facial expression recognition via deep action units graph network based on psychological mechanism.” *IEEE Transactions on Cognitive and Developmental Systems*, 12(2), 311–322.
- Loey, M., Manogaran, G., Taha, M. H. N. and Khalifa, N. E. M. (2021). “A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the covid-19 pandemic.” *Measurement*, 167, 108288–108298.
- Lopes, A. T., de Aguiar, E., De Souza, A. F. and Oliveira-Santos, T. (2017). “Facial expression recognition with convolutional neural networks: coping with few data and the training sample order.” *Pattern Recognition*, 61, 610–628.
- Lu, Z., Luo, Z., Zheng, H., Chen, J. and Li, W. (2015). “A delaunay-based temporal coding model for micro-expression recognition.” In *Computer Vision-ACCV 2014 Workshops*, Springer, 698–711.
- Luz, E., Silva, P., Silva, R., Silva, L., Guimarães, J., Miozzo, G., Moreira, G. and Menotti, D. (2021). “Towards an effective and efficient deep learning model for covid-19 patterns detection in x-ray images.” *Research on Biomedical Engineering*, 1–14.

## REFERENCES

---

- Ma, Y., Hao, Y., Chen, M., Chen, J., Lu, P. and Košir, A. (2019). “Audio-visual emotion fusion (avef): A deep efficient weighted approach.” *Information Fusion*, 46, 184–192.
- Majumder, A., Behera, L. and Subramanian, V. K. (2016). “Automatic facial expression recognition system using deep network-based data fusion.” *IEEE Transactions on Cybernetics*, 48(1), 103–114.
- Malmasi, S. and Dras, M. (2018). “Native language identification with classifier stacking and ensembles.” *Computational Linguistics*, 44(3), 403–446.
- Mare, T., Duta, G., Georgescu, M.-I., Sandru, A., Alexe, B., Popescu, M. and Ionescu, R. T. (2021). “A realistic approach to generate masked faces applied on two novel masked face recognition data sets.” *arXiv preprint arXiv:2109.01745*.
- Martinez, B., Valstar, M. F., Jiang, B. and Pantic, M. (2017). “Automatic analysis of facial actions: A survey.” *IEEE Transactions on Affective Computing*, 10(3), 325–347.
- Matias, R., Cohn, J. F. and Ross, S. (1989). “A comparison of two systems that code infant affective expression.” *Developmental Psychology*, 25(4), 483–489.
- Mavadati, M., Sanger, P. and Mahoor, M. H. (2016). “Extended disfa dataset: Investigating posed and spontaneous facial expressions.” In *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, IEEE, 1–8.
- Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P. and Cohn, J. F. (2013). “Disfa: A spontaneous facial action intensity database.” *IEEE Transactions on Affective Computing*, 4(2), 151–160.
- McCallum, A., Nigam, K. et al. (1998). “A comparison of event models for naive bayes text classification.” In *AAAI-98 workshop on learning for text categorization*, 41–48.
- Michel, P. and El Kaliouby, R. (2003). “Real time facial expression recognition in video using support vector machines.” In *5th International Conference on Multimodal Interfaces*, ACM, 258–264.
- Mihalcea, R. (2002). “Classifier stacking and voting for text filtering.” In *Text Retrieval Conference (TREC)*, 696–701.
- Min, R., Hadid, A. and Dugelay, J.-L. (2011). “Improving the recognition of faces occluded by facial accessories.” In *International Conference on Automatic Face & Gesture Recognition (FG)*, IEEE, 442–447.
- Min, R., Hadid, A. and Dugelay, J.-L. (2014). “Efficient detection of occlusion prior to robust face recognition.” *The Scientific World Journal*, 2014, 1–10.
- Minaee, S., Abdolrashidi, A., Su, H., Bennamoun, M. and Zhang, D. (2019). “Biometrics recognition using deep learning: A survey.” *arXiv preprint arXiv:1912.00271*.
- Mollahosseini, A., Chan, D. and Mahoor, M. H. (2016). “Going deeper in facial expression recognition using deep neural networks.” In *Winter conference on applications of computer vision (WACV)*, IEEE, 1–10.
- Mollahosseini, A., Hasani, B. and Mahoor, M. H. (2017). “Affectnet: A database for facial expression, valence, and arousal computing in the wild.” *IEEE Transactions on Affective Computing*, 10(1), 18–31.

- Munasinghe, M. (2018). “Facial expression recognition using facial landmarks and random forest classifier.” In *17th International Conference on Computer and Information Science (ICIS)*, IEEE, 423–427.
- Nair, V. and Hinton, G. E. (2010). “Rectified linear units improve restricted boltzmann machines.” In *27th International Conference on Machine Learning (ICML-10)*, ACM, 807–814.
- Namba, S., Makihara, S., Kabir, R. S., Miyatani, M. and Nakao, T. (2017). “Spontaneous facial expressions are different from posed facial expressions: morphological properties and dynamic sequences.” *Current Psychology*, 36(3), 593–605.
- Nan, F., Jing, W., Tian, F., Zhang, J., Chao, K.-M., Hong, Z. and Zheng, Q. (2022). “Feature super-resolution based facial expression recognition for multi-scale low-resolution images.” *Knowledge-Based Systems*, 236, 107678–107685.
- Navarathna, R., Carr, P., Lucey, P. and Matthews, I. (2017). “Estimating audience engagement to predict movie ratings.” *IEEE Transactions on Affective Computing*, 10(1), 48–59.
- Nazeri, K., Ng, E., Joseph, T., Qureshi, F. Z. and Ebrahimi, M. (2019). “Edgeconnect: Generative image inpainting with adversarial edge learning.” *arXiv preprint arXiv:1901.00212*.
- Nguyen, D. H., Kim, S., Lee, G.-S., Yang, H.-J., Na, I.-S. and Kim, S. H. “Facial expression recognition using a temporal ensemble of multi-level convolutional neural networks.” *IEEE Transactions on Affective Computing*, 13(1), 226–237.
- Niu, B., Gao, Z. and Guo, B. (2021). “Facial expression recognition with lbp and orb features.” *Computational Intelligence and Neuroscience*, 2021, 1–10.
- Niu, X., Han, H., Zeng, J., Sun, X., Shan, S., Huang, Y., Yang, S. and Chen, X. (2018). “Automatic engagement prediction with gap feature.” In *20th International Conference on Multimodal Interaction*, ACM, 599–603.
- Nkomo, L. M., Daniel, B. K. and Butson, R. J. (2021). “Synthesis of student engagement with digital technologies: a systematic review of the literature.” *International Journal of Educational Technology in Higher Education*, 18(1), 1–26.
- Nonis, F., Dagnes, N., Marcolin, F. and Vezzetti, E. (2019). “3d approaches and challenges in facial expression recognition algorithms—a literature review.” *Applied Sciences*, 9(18), 3904–3936.
- Oh, Y.-H., See, J., Le Ngo, A. C., Phan, R. C.-W. and Baskaran, V. M. (2018). “A survey of automatic facial micro-expression analysis: databases, methods, and challenges.” *Frontiers in psychology*, 9, 1128–1148.
- Ojala, T., Pietikäinen, M. and Harwood, D. (1996). “A comparative study of texture measures with classification based on featured distributions.” *Pattern Recognition*, 29(1), 51–59.
- Opitz, M., Waltner, G., Poier, G., Possegger, H. and Bischof, H. (2016). “Grid loss: Detecting occluded faces.” In *14th European Conference on Computer Vision (ECCV)*, Springer, 386–402.
- Pan, X., Ying, G., Chen, G., Li, H. and Li, W. (2019). “A deep spatial and temporal aggregation framework for video-based facial expression recognition.” *IEEE Access*, 7, 48807–48815.

## REFERENCES

---

- Peng, M., Wang, C., Bi, T., Shi, Y., Zhou, X. and Chen, T. (2019). “A novel apex-time network for cross-dataset micro-expression recognition.” In *8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 1–6.
- Pengcheng, W., Wang, Z., Ji, Z., Liu, X., Yang, S. and Wu, Z. (2020). “Tal emotionet challenge 2020 rethinking the model chosen problem in multi-task learning.” In *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, IEEE, 412–413.
- Pfister, T., Li, X., Zhao, G. and Pietikäinen, M. (2011). “Recognising spontaneous facial micro-expressions.” In *International Conference on Computer Vision*, IEEE, 1449–1456.
- Pham, T. T. D., Kim, S., Lu, Y., Jung, S.-W. and Won, C.-S. (2019). “Facial action units-based image retrieval for facial expression recognition.” *IEEE Access*, 7, 5200–5207.
- Picard, R. W., Vyzas, E. and Healey, J. (2001). “Toward machine emotional intelligence: Analysis of affective physiological state.” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (10), 1175–1191.
- Polikovskiy, S., Kameda, Y. and Ohta, Y. (2009). “Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor.” In *3rd International Conference on Crime Detection and Prevention (ICDP)*, IEEE, 1–6.
- Posner, J., Russell, J. A. and Peterson, B. S. (2005). “The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology.” *Development and psychopathology*, 17(3), 715–734.
- Pramerdorfer, C. and Kampel, M. (2016). “Facial expression recognition using convolutional neural networks: state of the art.” *arXiv preprint arXiv:1612.02903*.
- Qu, F., Wang, S.-J., Yan, W.-J., Li, H., Wu, S. and Fu, X. (2017). “CAS (ME)<sup>2</sup>: A database for spontaneous macro-expression and micro-expression spotting and recognition.” *IEEE Transactions on Affective Computing*, 9(4), 424–436.
- Rao, Q., Qu, X., Mao, Q. and Zhan, Y. (2015). “Multi-pose facial expression recognition based on surf boosting.” In *International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 630–635.
- Rao, R. S., Vaishnavi, T. and Pais, A. R. (2019). “Phishdump: A multi-model ensemble based technique for the detection of phishing sites in mobile devices.” *Pervasive and Mobile Computing*, 60, 101084–101098.
- Rashid, T. A. (2016). “Convolutional neural networks based method for improving facial expression recognition.” In *International Symposium on Intelligent Systems Technologies and Applications*, Springer, 73–84.
- Reddy, G. V., Savarni, C. D. and Mukherjee, S. (2020). “Facial expression recognition in the wild, by fusion of deep learnt and hand-crafted features.” *Cognitive Systems Research*, 62, 23–34.
- Renda, A., Barsacchi, M., Bechini, A. and Marcelloni, F. (2019). “Comparing ensemble strategies for deep learning: An application to facial expression recognition.” *Expert Systems with Applications*, 136, 1–11.
- Riaz, M. N., Shen, Y., Sohail, M. and Guo, M. (2020). “exnet: An efficient approach for emotion recognition in the wild.” *Sensors*, 20(4), 1087–1103.

- Rinn, W. E. (1984). “The neuropsychology of facial expression: a review of the neurological and psychological mechanisms for producing facial expressions.” *Psychological bulletin*, 95(1), 52–77.
- Ruan, D., Yan, Y., Lai, S., Chai, Z., Shen, C. and Wang, H. (2021). “Feature decomposition and reconstruction learning for effective facial expression recognition.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 7660–7669.
- Russell, J. A. (1980). “A circumplex model of affect.” *Journal of personality and social psychology*, 39(6), 1161–1178.
- Russell, J. A. (1991). “Culture and the categorization of emotions.” *Psychological bulletin*, 110(3), 426–450.
- Russell, J. A. (1994). “Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies.” *Psychological bulletin*, 115(1), 102–141.
- Rymarczyk, K., Żurawski, Ł., Jankowiak-Siuda, K. and Szatkowska, I. (2016). “Emotional empathy and facial mimicry for static and dynamic facial expressions of fear and disgust.” *Frontiers in psychology*, 7, 1853–1863.
- Safavian, S. R. and Landgrebe, D. (1991). “A survey of decision tree classifier methodology.” *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3), 660–674.
- Sakkis, G., Androustopoulos, I., Paliouras, G., Karkaletsis, V., Spyropoulos, C. D. and Stamatopoulos, P. (2001). “Stacking classifiers for anti-spam filtering of e-mail.” *arXiv preprint cs/0106040*.
- Samal, A. and Iyengar, P. A. (1992). “Automatic recognition and analysis of human faces and facial expressions: A survey.” *Pattern recognition*, 25(1), 65–77.
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.-C. (2018). “Mobilenetv2: Inverted residuals and linear bottlenecks.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 4510–4520.
- Saurav, S., Gidde, P., Saini, R. and Singh, S. (2022). “Dual integrated convolutional neural network for real-time facial expression recognition in the wild.” *The Visual Computer*, 38(3), 1083–1096.
- Sebe, N., Cohen, I., Gevers, T. and Huang, T. S. (2005). “Multimodal approaches for emotion recognition: a survey.” In *Internet Imaging VI*, SPIE, 56–68.
- Sebe, N., Lew, M. S., Sun, Y., Cohen, I., Gevers, T. and Huang, T. S. (2007). “Authentic facial expression analysis.” *Image and Vision Computing*, 25(12), 1856–1863.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017). “Grad-cam: Visual explanations from deep networks via gradient-based localization.” In *International Conference on Computer Vision*, IEEE, 618–626.
- Senechal, T., Bailly, K. and Prevost, L. (2014). “Impact of action unit detection in automatic emotion recognition.” *Pattern Analysis and Applications*, 17(1), 51–67.
- Sepas-Moghaddam, A., Etemad, A., Pereira, F. and Correia, P. L. (2021). “Capsfield: Light field-based face and expression recognition in the wild using capsule routing.” *arXiv preprint arXiv:2101.03503*.

## REFERENCES

---

- Shahar, H. and Hel-Or, H. (2019). “Micro expression classification using facial color and deep learning methods.” In *International Conference on Computer Vision Workshops*, 1–8.
- Shao, J. and Qian, Y. (2019). “Three convolutional neural network models for facial expression recognition in the wild.” *Neurocomputing*, 355, 82–92.
- Shen, J., Yang, H., Li, J. and Cheng, Z. (2021). “Assessing learning engagement based on facial expression recognition in mooc’s scenario.” *Multimedia Systems*, 1–10.
- Shokrani, S., Moallem, P. and Habibi, M. (2014). “Facial emotion recognition method based on pyramid histogram of oriented gradient over three direction of head.” In *4th International Conference on Computer and Knowledge Engineering (ICCKE)*, IEEE, 215–220.
- Shreve, M., Godavarthy, S., Goldgof, D. and Sarkar, S. (2011). “Macro-and micro-expression spotting in long videos using spatio-temporal strain.” In *International Conference on Automatic Face & Gesture Recognition (FG)*, IEEE, 51–56.
- Sifre, L. and Mallat, S. (2014). “Rigid-motion scattering for texture classification.” *arXiv preprint arXiv:1403.1687*.
- Sinatra, G. M., Heddy, B. C. and Lombardi, D. (2015). “The challenges of defining and measuring student engagement in science.” *Educational psychologist*, 50(1), 1–13.
- Song, L., Gong, D., Li, Z., Liu, C. and Liu, W. (2019). “Occlusion robust face recognition based on mask learning with pairwise differential siamese network.” In *International Conference on Computer Vision*, IEEE, 773–782.
- Sultan Zia, M., Hussain, M. and Arfan Jaffar, M. (2018). “A novel spontaneous facial expression recognition using dynamically weighted majority voting based ensemble classifier.” *Multimedia Tools and Applications*, 77, 25537–25567.
- Sun, B., Cao, S., He, J. and Yu, L. (2018). “Affect recognition from facial movements and body gestures by hierarchical deep spatio-temporal features and fusion strategy.” *Neural Networks*, 105, 36–51.
- Sun, N., Li, Q., Huan, R., Liu, J. and Han, G. (2019). “Deep spatial-temporal feature fusion for facial expression recognition in static images.” *Pattern Recognition Letters*, 119, 49–61.
- Sun, W., Zhao, H. and Jin, Z. (2017). “An efficient unconstrained facial expression recognition algorithm based on stack binarized auto-encoders and binarized neural networks.” *Neurocomputing*, 267, 385–395.
- Sun, X. and Xia, Pingping, S. L. (2020). “A roi-guided deep architecture for robust facial expressions recognition.” *Information Sciences*, 522, 35–48.
- Takalkar, M. A., Thuseethan, S., Rajasegarar, S., Chaczko, Z., Xu, M. and Yearwood, J. (2021). “Lgattnet: Automatic micro-expression detection using dual-stream local and global attentions.” *Knowledge-Based Systems*, 212, 106566–106575.
- Tan, M. and Le, Q. V. (2019). “Efficientnet: Rethinking model scaling for convolutional neural networks.” *arXiv preprint arXiv:1905.11946*.
- Tang, J., Alelyani, S. and Liu., H. (2015). “Data classification: Algorithms and applications.” In *Data Mining and Knowledge Discovery Series*, CRC Press, 498–500.



- Tang, Y., Zhang, X. M. and Wang, H. (2018). “Geometric-convolutional feature fusion based on learning propagation for facial expression recognition.” *IEEE Access*, 6, 42532–42540.
- Thi Thu Nguyen, N., Thi Thu Nguyen, D. and The Pham, B. (2021). “Micro-expression recognition based on the fusion between optical flow and dynamic image.” In *5th International Conference on Machine Learning and Soft Computing*, ACM, 115–120.
- Tian, Y.-l., Kanada, T. and Cohn, J. F. (2000). “Recognizing upper face action units for facial expression analysis.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, IEEE, 294–301.
- Tian, Y.-L., Kanade, T. and Cohn, J. F. (2005). “Facial expression analysis.” In *Handbook of face recognition*, Springer, 247–275.
- Tong, Z., Zheng, W., Cui, Z., Zong, Y., Yan, J. and Yan, K. (2016). “A deep neural network-driven feature learning method for multi-view facial expression recognition.” *IEEE Transactions on Multimedia*, 18(12), 2528–2536.
- Valstar, M. and Pantic, M. (May 2010). “Induced disgust, happiness and surprise: an addition to the mmi facial expression database.” In *3rd International Workshop on EMOTION*, ACM, 65–70.
- Valstar, M. F., Mehu, M., Jiang, B., Pantic, M. and Scherer, K. (2012). “Meta-analysis of the first facial expression recognition challenge.” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(4), 966–979.
- Venkatesh, S. and Koolagudi, S. G. (2022). “Device robust acoustic scene classification using adaptive noise reduction and convolutional recurrent attention neural network.” In *24th International Conference on Speech and Computer (SPECOM)*, Springer, 688–699.
- Verma, M., Vipparthi, S. K., Singh, G. and Murala, S. (2019). “Learnet: Dynamic imaging network for micro expression recognition.” *IEEE Transactions on Image Processing*, 29, 1618–1627.
- Viola, P. and Jones, M. (2001). “Rapid object detection using a boosted cascade of simple features.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, IEEE, 511–518.
- Viola, P. and Jones, M. J. (2004). “Robust real-time face detection.” *International journal of computer vision*, 57(2), 137–154.
- Vo, T.-H., Lee, G.-S., Yang, H.-J. and Kim, S.-H. (2020). “Pyramid with super resolution for in-the-wild facial expression recognition.” *IEEE Access*, 8, 131988–132001.
- Wan, W. and Chen, J. (2017). “Occlusion robust face recognition based on mask learning.” In *International Conference on Image Processing (ICIP)*, IEEE, 3795–3799.
- Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X. and Tang, X. (2017). “Residual attention network for image classification.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 3156–3164.
- Wang, K., Peng, X., Yang, J., Lu, S. and Qiao, Y. (2020a). “Suppressing uncertainties for large-scale facial expression recognition.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 6897–6906.

## REFERENCES

---

- Wang, K., Peng, X., Yang, J., Meng, D. and Qiao, Y. (2020b). “Region attention networks for pose and occlusion robust facial expression recognition.” *IEEE Transactions on Image Processing*, 29, 4057–4069.
- Wang, S., Wu, C., He, M., Wang, J. and Ji, Q. (2015). “Posed and spontaneous expression recognition through modeling their spatial patterns.” *Machine Vision and Applications*, 26(2), 219–231.
- Wang, S.-J., Yan, W.-J., Li, X., Zhao, G. and Fu, X. (2014). “Micro-expression recognition using dynamic textures on tensor independent color space.” In *22nd International Conference on Pattern Recognition*, IEEE, 4678–4683.
- Wang, Z., Huang, B., Wang, G., Yi, P. and Jiang, K. (2023). “Masked face recognition dataset and application.” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1–7.
- Weber, R., Soladié, C. and Séguier, R. (2018). “A survey on databases for facial expression analysis.” In *13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*, Springer, 73–84.
- Wen, G., Hou, Z., Li, H., Li, D., Jiang, L. and Xun, E. (2017). “Ensemble of deep neural networks with probability-based fusion for facial expression recognition.” *Cognitive Computation*, 9(5), 597–610.
- Wen, Y., Zhang, K., Li, Z. and Qiao, Y. (2016). “A discriminative feature learning approach for deep face recognition.” In *14th European Conference on Computer Vision (ECCV)*, Springer, 499–515.
- Wenyun, S., Zhao, H. and Jin, Z. (2018). “A complementary facial representation extracting method based on deep learning.” *Neurocomputing*, 306, 246–259.
- Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A. and Movellan, J. R. (2014). “The faces of engagement: Automatic recognition of student engagement from facial expressions.” *IEEE Transactions on Affective Computing*, 5(1), 86–98.
- Wickramaratne, S. D. and Mahmud, M. S. (2020). “Bi-directional gated recurrent unit based ensemble model for the early detection of sepsis.” In *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 70–73.
- Wu, B.-F. and Lin, C.-H. (2018). “Adaptive feature mapping for customizing deep learning based facial expression recognition model.” *IEEE access*, 6, 12451–12461.
- Xia, Y., Zhang, B. and Coenen, F. (2015). “Face occlusion detection based on multi-task convolution neural network.” In *12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, IEEE, 375–379.
- Xiaohua, W., Muzi, P., Lijuan, P., Min, H., Chunhua, J. and Fuji, R. (2019). “Two-level attention with two-stage multi-task learning for facial emotion recognition.” *Journal of Visual Communication and Image Representation*, 62, 217–225.
- Xie, H.-X., Lo, L., Shuai, H.-H. and Cheng, W.-H. (2020). “Au-assisted graph attention convolutional network for micro-expression recognition.” In *28th International Conference on Multimedia*, ACM, 2871–2880.
- Xie, S. and Hu, H. (2018). “Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks.” *IEEE Transactions on Multimedia*, 21(1), 211–220.

- Xie, S., Hu, H. and Wu, Y. (2019). “Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition.” *Pattern Recognition*, 92, 177–191.
- Yaddaden, Y., Adda, M., Bouzouane, A., Gaboury, S. and Bouchard, B. (2018). “Hybrid-based facial expression recognition approach for human-computer interaction.” In *20th International Workshop on Multimedia Signal Processing (MMSP)*, IEEE, 1–6.
- Yan, J., Zheng, W., Cui, Z., Tang, C., Zhang, T. and Zong, Y. (2018). “Multi-cue fusion for emotion recognition in the wild.” *Neurocomputing*, 309, 27–35.
- Yan, W.-J., Li, S., Que, C., Pei, J. and Deng, W. (2020). “Raf-au database: In-the-wild facial expressions with subjective emotion judgement and objective au annotations.” In *Asian Conference on Computer Vision*, Springer, 1–14.
- Yan, W.-J., Li, X., Wang, S.-J., Zhao, G., Liu, Y.-J., Chen, Y.-H. and Fu, X. (2014). “Casmie ii: An improved spontaneous micro-expression database and the baseline evaluation.” *PloS one*, 9(1), Art. No. e86041.
- Yang, H., Ciftci, U. and Yin, L. (2018). “Facial expression recognition by de-expression residue learning.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2168–2177.
- Yanliang, Z., Liu, Y., Li, G. and Peng, H. (2021). “Subspace learning and joint distribution adaptation for unsupervised cross-database microexpression recognition.” *Mobile Information Systems*, 2021, 1–8.
- Yi, D., Lei, Z., Liao, S. and Li, S. Z. (2014). “Learning face representation from scratch.” *arXiv preprint arXiv:1411.7923*.
- Yu, Z. and Zhang, C. (2015). “Image based static facial expression recognition with multiple deep network learning.” In *ACM International Conference on Multimodal Interaction*, 435–442.
- Zeng, D., Veldhuis, R. and Spreuwers, L. (2021). “A survey of face recognition techniques under occlusion.” *IET biometrics*, 10(6), 581–606.
- Zhang, K., Huang, Y., Du, Y. and Wang, L. (2017). “Facial expression recognition based on deep evolutionary spatial-temporal networks.” *IEEE Transactions on Image Processing*, 26(9), 4193–4203.
- Zhang, K., Zhang, Z., Li, Z. and Qiao, Y. (2016). “Joint face detection and alignment using multitask cascaded convolutional networks.” *IEEE signal processing letters*, 23(10), 1499–1503.
- Zhang, T., Zhang, X., Shi, J. and Wei, S. (2019). “Depthwise separable convolution neural network for high-speed sar ship detection.” *Remote Sensing*, 11(21), 2483–2519.
- Zhang, Y., Liu, Y. and Wang, H. (2021). “Cross-database micro-expression recognition exploiting intradomain structure.” *Journal of Healthcare Engineering*, 2021, 1–9.
- Zhang, Z. (2018). “Improved adam optimizer for deep neural networks.” In *26th International Symposium on Quality of Service (IWQoS)*, IEEE, 1–2.
- Zhang, Z., Luo, P., Loy, C. C. and Tang, X. (2018). “From facial expression recognition to interpersonal relation prediction.” *International Journal of Computer Vision*, 126(5), 550–569.

## REFERENCES

---

- Zhao, G., Huang, X., Taini, M., Li, S. Z. and Pietikäinen, M. (2011). “Facial expression recognition from near-infrared videos.” *Image and Vision Computing*, 29(9), 607–619.
- Zhao, G., Yang, H. and Yu, M. (2020). “Expression recognition method based on a lightweight convolutional neural network.” *IEEE Access*, 8, 38528–38537.
- Zhao, X., Liang, X., Liu, L., Li, T., Han, Y., Vasconcelos, N. and Yan, S. (2016). “Peak-piloted deep network for facial expression recognition.” In *European Conference on Computer Vision*, Springer, 425–442.
- Zhao, X., Zou, J., Li, H., Dellandréa, E., Kakadiaris, I. A. and Chen, L. (2015). “Automatic 2.5-d facial landmarking and emotion annotation for social interaction assistance.” *IEEE Transactions on Cybernetics*, 46(9), 2042–2055.
- Zhi, R., Liu, M. and Zhang, D. (2020). “A comprehensive survey on automatic facial action unit analysis.” *The Visual Computer*, 36(5), 1067–1093.
- Zhi, R., Zhou, C., Li, T., Liu, S. and Jin, Y. (2021). “Action unit analysis enhanced facial expression recognition by deep neural network evolution.” *Neurocomputing*, 425, 135–148.
- Zhong, L., Liu, Q., Yang, P., Huang, J. and Metaxas, D. N. (2014). “Learning multiscale active facial patches for expression analysis.” *IEEE Transactions on Cybernetics*, 45(8), 1499–1510.
- Zhu, B., Lan, X., Guo, X., Barner, K. E. and Boncelet, C. (2020a). “Multi-rate attention based gru model for engagement prediction.” In *International Conference on Multimodal Interaction*, ACM, 841–848.
- Zhu, J., Luo, B., Zhao, S., Ying, S., Zhao, X. and Gao, Y. (2020b). “Iexpressnet: Facial expression recognition with incremental classes.” In *28th International Conference on Multimedia*, ACM, 2899–2908.
- Zhu, Q., Mao, Q., Jia, H., Noi, O. E. N. and Tu, J. (2022). “Convolutional relation network for facial expression recognition in the wild with few-shot learning.” *Expert Systems with Applications*, 189, 116046–116054.
- Zia, M. S., Hussain, M. and Jaffar, M. A. (2018). “A novel spontaneous facial expression recognition using dynamically weighted majority voting based ensemble classifier.” *Multimedia Tools and Applications*, 77(19), 25537–25567.
- Zong, Y., Zheng, W., Huang, X., Shi, J., Cui, Z. and Zhao, G. (2018). “Domain regeneration for cross-database micro-expression recognition.” *IEEE Transactions on Image Processing*, 27(5), 2484–2498.

# PUBLICATIONS

## JOURNAL PAPERS

1. Rashmi Adyapady R., Annappa B. (2022). A Comprehensive Review of Facial Expression Recognition Techniques. *Multimedia Systems*, Springer, 29(1), 73–103.  
(DOI:<https://doi.org/10.1007/s00530-022-00984-w>)
2. Rashmi Adyapady R., Annappa B. (2022). Micro Expression Recognition Using Delaunay Triangulation and Voronoi Tessellation. *IETE Journal of Research*, Taylor & Francis, 1–17.  
(DOI:<https://doi.org/10.1080/03772063.2022.2068680>)
3. Rashmi Adyapady R., Annappa B. (2022). An Ensemble Approach using a Frequency-Based and Stacking Classifiers for Effective Facial Expression Recognition. *Multimedia Tools and Applications*, Springer, 82(10). 14689–14712.  
(DOI:<https://doi.org/10.1007/s11042-022-13940-7>)

## CONFERENCE PAPERS

1. Rashmi Adyapady R., Annappa B. (2022). Learning Engagement Assessment in MOOC Scenario. In *Proceedings of the 8th IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, IEEE, pp. 1–6.  
(DOI:<https://doi.org/10.1109/CONECCT55679.2022.9865699>)

## REFERENCES

---

2. Rashmi Adyapady R., Annappa B. (2023). An Xception Model with Residual Attention Mechanism for Facial Occlusion Detection. In Proceedings of the 8th International Conference for Convergence in Technology (I2CT), IEEE, pp. 1–6. (DOI:<https://doi.org/10.1109/I2CT57861.2023.10126182>)

## BIODATA

**Name:** RASHMI ADYAPADY R.  
**Date of Birth:** 5<sup>th</sup> JUNE 1989  
**Gender:** Female  
**Marital Status:** Married  
**Father's Name:** Raghava A.  
**Mother's Name:** Poovamma  
**Address:** "City Garden" Apartment  
Flat No. 103,  
Kalappayya Fisheries Road,  
City Garden, Iddya, Surathkal  
Mangaluru-575014  
Dakshina Kannada, Karnataka, India  
**E-mail:** [rashmiadyapady5@gmail.com](mailto:rashmiadyapady5@gmail.com)  
**Mobile:** +91 9900302491  
**Qualification:** B.E in Computer Science & Engineering  
M.Tech in Computer Science & Engineering  
(NMAM Institute of Technology (NMAMIT,  
Nitte))  
**Areas of Interest:** Image Processing, Machine Learning,  
Deep Learning